Charles University

Faculty of Science

Study programme: Special Chemical and Biological Programmes

Branch of study: Molecular Biology and Biochemistry of Organisms

Daniel Žucha

The importance and role of reverse transcriptases in gene expression analysis

Význam a rola reverzných transkriptáz pri analýze génovej expresie

Bachelor's thesis

Supervisor: Ing. Lukáš Valihrach, Ph.D.

Prague, 2018

Acknowledgement

Having the opportunity, I would like to express gratitude to my supervisor Ing. Lukáš Valihrach, Ph.D. for his help and support throughout the studies. Similarly, I would like to thank my family and friends for all the kindness and help they provided.

Prehlásenie

Prehlasujem, že som záverečnú prácu spracoval samostatne a že som uviedol všetky použité infromačné zdroje a literatúru. Táto práca ani jej podstatná časť nebola predložená k získaniu iného alebo rovnakého akademického titulu.

V Prahe, 30.4.2018

Podpis

Abstract

The continuously advancing field of gene expression analysis enables the evaluation of even the slightest changes that occur in the cell transcriptome. In order to ensure accuracy of the observed biological variances, it is fundamentally important to be aware of the possible biases introduced during sample processing. In gene expression research, the methods of reverse transcription–quantitative PCR (RT–qPCR) and RNA-Sequencing (RNA-Seq) are often the primary choice, mostly because of their high precision and reproducibility. Since these both methods require DNA template, they are coupled with the same initial step - reverse transcription (RT), a reaction producing DNA complementary to its RNA template. It is well known that RT introduces bias. As a result, it is therefore of importance to thoroughly evaluate the effects of these biases. One such annotated source of artifacts is the reverse transcriptase (RTase) itself. However, it has been shown that the enzyme does not account for most of the variance alone. Surprisingly, choice of primers or RNA template may influence the reaction outcome even more than the bias introduced from the enzyme. This is especially the case with recent advances in protein engineering. Production of highly efficient RTases may pronounce the variation originating from other reaction components. This thesis is focused on the RTase characteristics and factors influencing RT reaction.

Keywords: reverse transcription, gene expression, reverse transcription–quantitative PCR, RNA-Sequencing, oligonucleotide

Abstrakt

Neustále napredovanie v oblasti analýzy génovej expresie umožňuje zaznamenať aj tie najmenšie zmeny, ktoré sa odohrávajú v transkriptóme bunky. Pre zaistenie správnosti pozorovanej biologickej odlišnosti, je dôležité byť si vedomý možných chýb, ktoré sú dôsledkom spracovania vzoriek. Vo výskume génovej expresie sú často metódy reverznej transkripcie–kvantitatívnej PCR (RT–qPCR) a RNA sekvenovania (RNA-Seq) prvotnou voľbou, najmä z dôvodu ich vysokej presnosti a reprodukovateľnosti. Nakoľko obe tieto metódy potrebujú DNA ako templát, často im predchádza krok reverznej transkripcie (RT), reakcie syntetizujúcej DNA komplementárnu k RNA vláknu. Je známe, že RT sa do určitej miery vyznačuje chybovosťou, čo je práve podnetom na jej podrobné preskúmanie. Zdrojom artefaktov môže byť samotná reverzná transkriptáza (RTase), ale bolo preukázané, že enzým nie je jediným zdrojom týchto chýb. Výber stratégie primingu alebo RNA templát samotný môžu ovplyvniť výsledok reakcie ešte vo väčšej miere. Najmä v spojitosti so súčasným pokrokom v proteínovom inžinierstve, produkujúcom vysoko efektívne reverzné transkriptázy, sa variácia spôsobená ostatnými komponentmi reakcie ešte viac dostáva do popredia. Táto práca sa zameriava na charakteristiku reverznej transkriptázy a faktorov ovplyvňujúcich reverznú transkripciu.

Kľúčové slová: reverzná transkripcia, génová expresia, reverzná transkripcia–kvantitatívna PCR, RNA sekvenovanie, oligonukleotid

Abbreviations

AMV = **A**vian **m**yeloblastosis **v**irus

BrdU = 5-**br**omo**d**eoxy**u**ridine

cDNA = **c**omplementary **DNA**

Cq = **c**ycle of **q**uantification

CV = **c**oefficient of **v**ariation

dNTP = **d**eoxyribo**n**ucleotide **t**ri**p**hosphate

EC number = **E**nzyme **C**ommission **number**

FFPE = **f**ormalin-**f**ixed **p**araffin-**e**mbedded

GSP = **g**ene-**s**pecific **p**rimers

HIV = **H**uman **i**mmunodeficiency **v**irus

LINE = **l**ong **i**nterspersed **n**uclear **e**lements

MIQE = The **M**inimum **I**nformation for Publication of **Q**uantitative Real-Time PCR **E**xperiments

miRNA = **mi**cro**RNA**

MLV = **M**urine **l**eukemia **v**irus

MMLV = **M**oloney **m**urine **l**eukemia **v**irus

PCR = **p**olymerase **c**hain **r**eaction

qPCR = **q**uantitative real-time **PCR**

RNase = **r**ibo**n**ucle**ase**

RNA-Seq = **RNA-Seq**uencing

RSV = **R**ous **s**arcoma **v**irus

RT = **r**everse **t**ranscription

RTase = **r**everse **t**ranscript**ase**

RT–qPCR = **r**everse **t**ranscription–**q**uantitative **p**olymerase **c**hain **r**eaction

scRNA-Seq = **s**ingle-**c**ell **RNA-Seq**

SD = **s**tandard **d**eviation

SINE = **s**hort **i**nterspersed **n**uclear **e**lement

snoRNA = **s**mall **n**ucle**o**lar **RNA**

Tm = **m**elting **t**emperature

TSO = **t**emplate **s**witching **o**ligonucleotide

UMI = **u**nique **m**olecular **i**dentifiers

VSV = **V**esicular **s**tomatitis **v**irus

σ_e = posterior probability density of infinite populations SD parameter

Table of Contents

1 Introduction

Reverse transcriptase (RTase) is an enzyme capable of transcribing an RNA molecule into complementary DNA (cDNA), which is later used for downstream applications. The finding of such an enzyme was initially met with skepticism, as its existence was not implied by Crick's central dogma of molecular biology (Crick, 1958; Crick, 1970). Nevertheless, its discovery was presented upon simple, yet powerful claims (Baltimore, 1970; Temin & Mizutani, 1970) and a new era of molecular biology began.

Nowadays, reverse transcription (RT) is an unavoidable step to study the cell's decision making – the transcriptome. Since RT is often used in combination with methods of precise template quantification e.g. reverse transcription–quantitative PCR (RT–qPCR), the accuracy and reproducibility of RT are of the highest importance.

RTase, as the main component of this reaction, has been often reported to have a considerable impact on the outcome of the reaction (Ståhlberg *et al.*, 2004a; Levesque-Sergerie *et al.*, 2007; Lindén *et al.*, 2012; Bustin *et al.*, 2015). Additionally, parameters like choice of primers, samples or background RNA may exhibit a similar degree of influence on the reaction outcome (Zhang & Byrne, 1999; Lekanne Deprez *et al.*, 2002; Ståhlberg *et al.*, 2004b; Stangegaard *et al.*, 2006; Levesque-Sergerie *et al.*, 2007; Miranda & Steward, 2017). Therefore, a better understanding of the reaction's mechanism and its components may lead to further advances in gene expression analysis.

1.1 Aims

The main goal of this thesis is to review the current state of knowledge on the impact of the RTase and other reaction components on the outcome of RT reaction. Since the reaction is indispensably coupled with RT–qPCR and RNA-Seq, both methods are part of discussion.

2 History

The history of RTases is closely connected with the discoveries of viruses in chicken during the early twentieth century (Rous, 1911), which was later followed by viruses with similar properties in mammals (Perk & Moloney, 1966; Bittner, 1936). In the beginning, both classes were joined under a single name – oncoviruses (Valladares, 1960).

Howard Temin, who was the leading researcher in discovery of viral RTase, initiated his study of retroviruses through his observations of the Rous sarcoma virus (RSV). He remarked that different RSV strains induced different shapes in the infected host cells (Temin, 1960). Unsure of the complete mechanism,

he proposed two explanations for his findings: 1) virus provided its genetic information directly to the cell; 2) virus induced the tumorous response. His work, however, led him to believe that the first option was correct and the virus really donated some of its genetic information to the cell. Temin continued in his work, tracking the influence of various inhibitors on nucleic acid synthesis. Interestingly, he found that inhibitors of DNA-dependent RNA transcription blocked synthesis of RSV virions when added immediately after the infection (Temin, 1963; Temin 1964). On multiple occasions, Temin failed to persuade his fellow colleagues about RTase existence, despite none of his results were contradictory to the provirus theory. The general attitude however started to change, when Boettiger and Temin (1970) presented results of a study, where they used thymine substitute sensitive to light. When 5-bromodeoxyuridine (BrdU) was added immediately after the infection, longer exposures to light caused declines in virion formation. This became a proof that DNA plays a major role in the life cycle of sarcoma virus. The final observation of deoxyribonucleotide triphosphate incorporation by RNA-dependent DNA polymerase was published in June 1970 (Temin & Mizutani, 1970).

The next scientist who significantly contributed to the discovery of RTases was David Baltimore. His work in the field was initiated by exploration of the presence of an RNA polymerase in an enveloped virion of negative RNA strand virus – vesicular stomatitis virus (VSV). After the enzyme's presence was confirmed (Baltimore *et al.*, 1970), he changed his interest to RNA tumor viruses. Firstly, he tested whether Rauscher Murine leukemia virus (MLV) also had RNA polymerase activity but to no success. However, exchange of ribonucleotides for deoxyribonucleotides resulted in DNA synthesis in RNA tumor virus (Baltimore, 1970). The Baltimore group's following work utilized purified RTase to reverse transcribe eukaryotic mRNA (Verman *et al.*, 1974).

The simultaneous RTase discovery (Baltimore, 1970; Temin & Mizutani, 1970) by two independent research teams raised a lot of attention. Later work on the RTases elucidated the mechanisms of retroviral replication (Gilboa *et al.*, 1979), as well as discoveries of human pathologies, such as Human immunodeficiency virus (HIV) (Barré-Sinoussi *et al.*, 1983). Similarly, the discovery of proto-oncogene *c-src* made Oppermann *et al.* (1979) hypothesize that viral oncogenes might have developed from normal cellular genes. The importance of viruses for making up eukaryotic genomes was however discovered after genome sequencing took place. It was found that large parts of the eukaryotic genomes consist of genetic elements similar to retroviruses, e.g. long and short interspersed nuclear elements (LINEs and SINEs) (Lander *et al.*, 2001; Singer, 1982). Additionally, although RTase is known as a viral enzyme, it is also constantly functioning in humans. Since conventional DNA polymerases have problems with replication of chromosomal ends, telomerase enzyme reverse transcribes an RNA template contained within its subunit and synthesizes chromosome termini (Lingner *et al.*, 1997).

The simplicity of using RTase has led to rapid development of laboratory methods utilizing its unique properties. Many various methods were developed but just the combination of RT, PCR and fluorescence reporters enabled to study cell transcriptomics in real time (Gibson *et al.*, 1996). Currently, with the rise of next-generation sequencing techniques, gene expression can be studied on whole transcriptome level. RNA-Sequencing (RNA-Seq) allows for precise mapping of thousands of genes in a single run, making it currently the most powerful tool for transcriptome analysis (Nagalakshmi *et al.*, 2008; Mortazavi *et al.*, 2008).

3 Enzyme

RTases belong to a broader group of nucleotidyltransferases, carrying the Enzyme Commission number (EC number) 2.7.7.49 and official name of RNA-directed DNA polymerases. Although RTases are a broad category of enzymes, they share two common functions necessary for correct viral replication: 1) a DNA polymerase activity, with either RNA or DNA as a template, and 2) a ribonuclease H activity (RNase H), hydrolyzing RNA in RNA/DNA duplex (Coté & Roth, 2008).

3.1 Enzyme engineering

For *in vitro* purposes, mostly engineered versions of the enzyme are used because of their enhanced properties. These properties are achieved by specific mutations in the RTase structure. The focus of RTase engineering is usually for the improvement of the enzyme's: 1) thermostability, which allows for higher reaction temperatures, permitting the template's secondary structures to unfold (Arezi & Hogrefe, 2009); 2) processivity – ability to synthesize longer transcripts without releasing the substrate (Baranauskas *et al.*, 2012); 3) resistance to inhibitors (Arezi *et al.*, 2010); 4) fidelity – exactness of complementary nucleotide incorporation (Alvarez *et al.*, 2009); or 5) ability to reduce the activity of RNase H (Mizuno *et al.*, 2010).

3.2 Structure

Due to its medical importance, HIV-1 RTase is one of the most well-studied RTases, both structurally and functionally. However, HIV-1 RTase's lack of exonucleolytic proofreading activity and high error rate make it an unsuitable candidate for *in vitro* applications (Roberts *et al.*, 1988). Instead, engineered versions of Avian myeloblastosis virus (AMV) or Moloney murine leukemia virus (MMLV) RTases are often used.

MMLV RTase is a functional monomer of 671 amino acids that shares similar functionality as the 66-kDa subunit (p66) of the HIV-1 RTase but differs in its steric composition and sequence (Coté & Roth, 2008). MMLV RTase consists of 5 domains - fingers, thumb, palm, connection and RNase H domain (Figure 1A). All domains participate in the formation of the nucleic acid cleft (Das & Georgiadis, 2004).

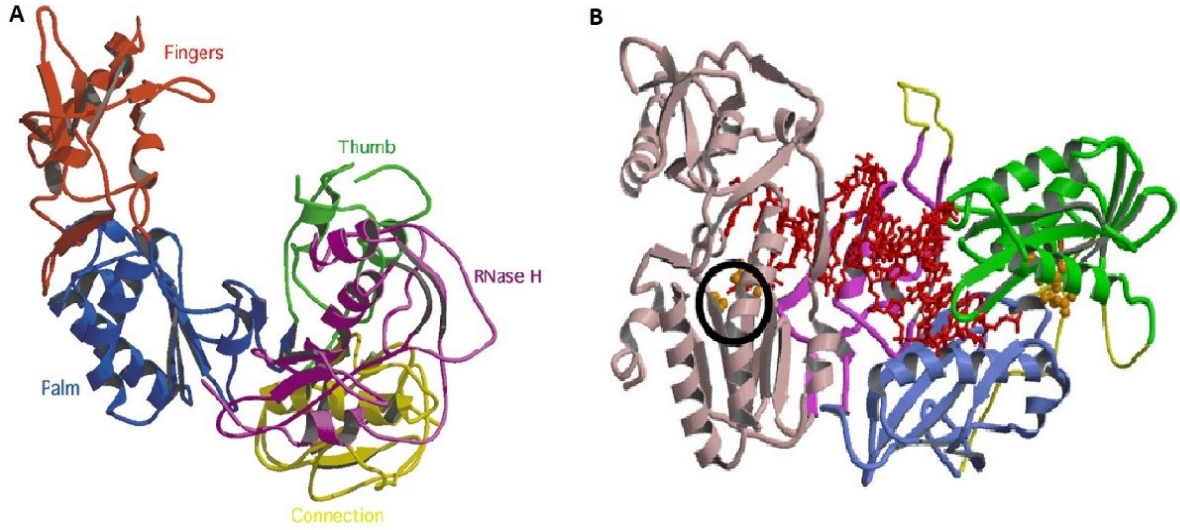

Figure 1: **The crystal structure of MMLV RTase with color-labeled domains.** (A) A ribbon diagram with fingers domain in red, palm domain in blue, thumb domain in green, connection domain in yellow and RNase H domain in purple (Das & Georgiadis, 2004). (B) An enzyme model bound with its dsDNA substrate (colored in red). The polymerase active site residues are colored in orange and encircled in black, while RNase H active residues are shown as orange dots only. Domain color coding is as follows: fingers and palm domains – pink; thumb – purple; connection – blue; RNase H domain – green (Coté & Roth, 2008).

The DNA polymerase active site is present at the N-terminus in the palm/fingers domain and consists of 3 aspartic acids residing on a β-sheet (Figure 1B). Two aspartates are part of the highly conserved retroviral RTase motif YXDD (Tyrosine – X – Aspartic acid – Aspartic acid), where X is an amino acid specific for each RTase. The polymerase active site binds two divalent ions, essential for catalysis, where Mg^{2+} is used in vivo, but Mn^{2+} is an option for in vitro reactions (Coté & Roth, 2008).

The RNase H active site is located in the enzyme's C-terminus and comprises of carboxylic acids – three aspartates and one glutamate. These amino acids are a part of the strictly conserved D-E-D motif (Aspartic acid – Glutamic acid – Aspartic acid). In general, RNase H domain presents a highly conserved protein fold across RNase H domains from divergent organisms, consisting of a mixed β sheet (three antiparallel and two parallel strands) and two to five α helices positioned in a shape of letter "H" (Nowotny *et al.*, 2005; Beilhartz & Götte, 2010). The active site of the RNase H domain lies ~ 60 Å from the polymerase active site, which corresponds to approximately 18 base pairs (Beilhartz & Götte, 2010).

3.3 DNA polymerization and RNase H cleavage

3.3.1 DNA polymerase

As with many DNA polymerases, RTase also requires a template and a primer. The mechanism is conserved among the myriad of the RTases, however, the structural and sequence differences may impact the details of the reaction (Das & Georgiadis, 2004).

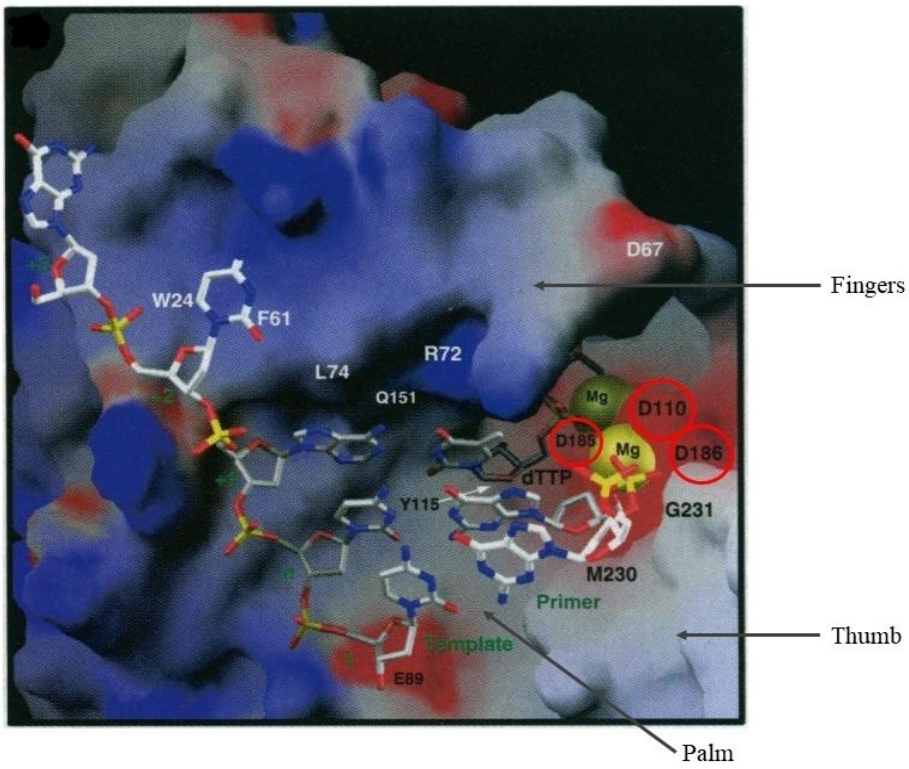

Figure 2: A view into the polymerase active site of HIV-1 RTase. The dNTP pocket is formed by domains as indicated with arrows: fingers (upper foreground), palm (middle background), thumb (lower right foreground). The template, primer and dNTP are presented in stick rendering. Protein surface is in continuous rendering. Divalent ions (Mg^{2+}) are presented as green and yellow spheres. Aspartic acids of the active site are encircled in red (adapted from Huang *et al.*, 1998).

The process of polymerization begins with the RTase binding the nucleic acid substrate, leading to a conformational change of thumb domain from "closed" to "open". It is known that RTase favors the binding of double-strand nucleic acids while positioning the primer's 3' end into the priming site (Meyer *et al.*, 2007). The priming site is located near the polymerase active site (Figure 2). "Open" conformation of the thumb domains allows for the initiation of the nucleotide incorporation. Incoming dNTP is bound in the nucleotide binding site, causing the fingers to enclose the dNTP (Huang *et al.*, 1998). This ensures the correct positioning of the 3' end of the primer, the dNTP's α-phosphate, and polymerase active site. After creation of the phosphodiester bond between the primer and the nucleotide (Figure 3), the fingers open, releasing generated

pyrophosphate (Malik *et al.*, 2017). In order to continue with the synthesis, the nucleic acid substrate must translocate, allowing another dNTP to be bound in the nucleotide binding site.

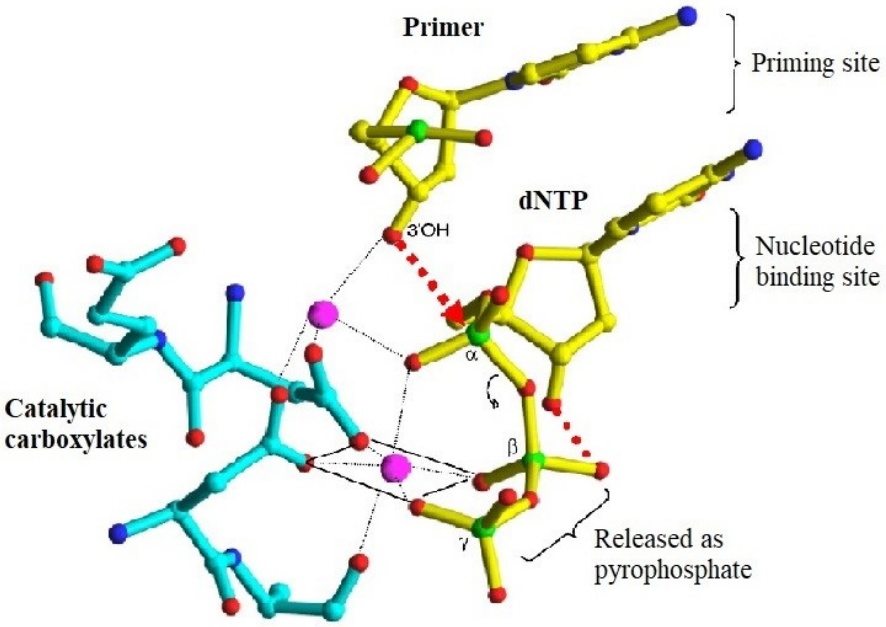

Figure 3: **The ternary complex of an enzyme/dNTP/primer in the polymerase active site.** The amino acids and nucleotides are presented in blue and yellow stick model, respectively. Divalent ions are purple spheres. Black dotted lines present possible interactions between molecules (Sarafianos *et al.*, 2009).

3.3.2 Ribonuclease H

RNA/DNA hybrid is recognized for its unique mixed composition of A and B conformation (Nowotny *et al.*, 2005), possibly due to the minor groove width of ~ 9-10 Å (Sarafianos *et al.*, 2001). The strictly conserved D-E-D motif utilizes two divalent ions (Mg^{2+}, Mn^{2+}) in the hydrolysis of the RNA substrate. The presence of 2'-hydroxyl group on the ribose of RNA strand makes this reaction RNA strand-specific (Nowotny *et al.*, 2005). A two-metal ion mechanism utilizes a water molecule as a nucleophile in order to enclose the distance between the metal ions. The nucleophile attack on the scissile bond is carried out by the water molecule. Both metal ions have a role in the stabilization of the transition state. The product of this reaction is 3'- hydroxyl and 5'-phosphate group. Metal ions are recovered, and the reaction can be repeated (Figure 4).

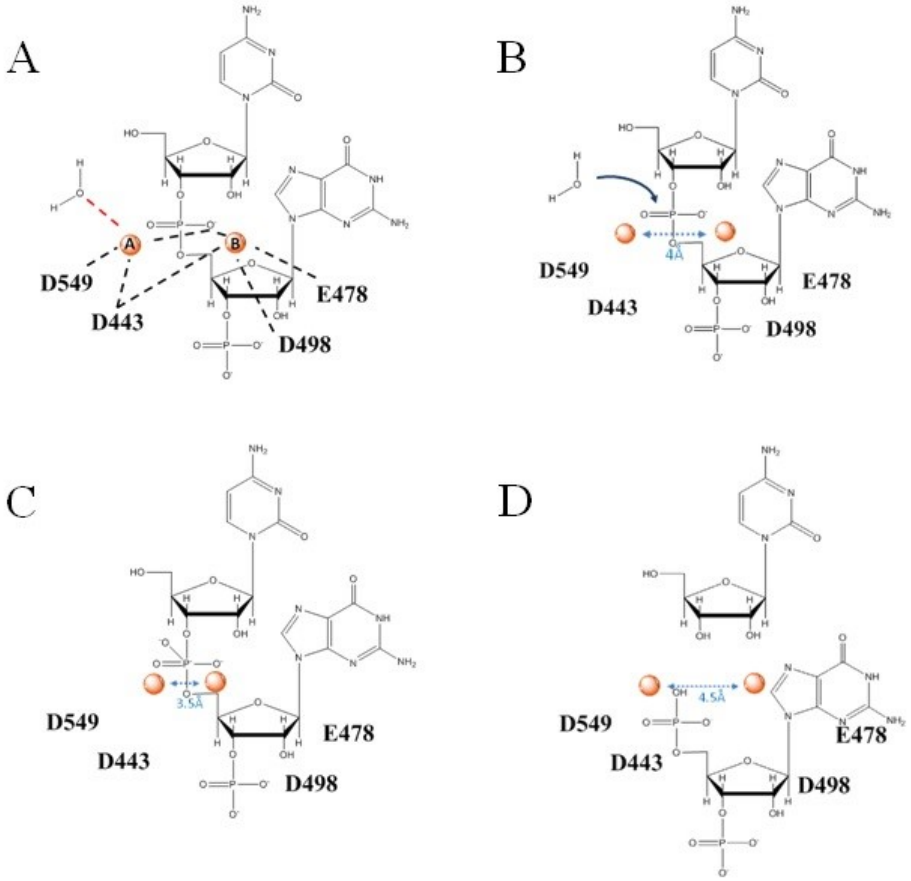

Figure 4: **A two-metal ion mechanism of HIV-1 RNase H.** Metal ions are presented as orange spheres. (A) Active site carboxylic residues coordinate the metal ions. Activation of water molecule is represented by red dashed line. (B) The nucleophilic attack (blue arrow) is carried out by water molecule. (C) Nucleophile is brought closer to phosphate by metal ions moving closer to each other. (D) Product of the reaction is released: ribonucleotide with 3'- hydroxyl group and 5'- phosphate group. The distance between metal ions is recovered (Beilhartz & Götte, 2010).

3.4 Additional enzyme properties

3.4.1 Terminal transferase and template switching activity

Terminal transferase activity of an enzyme is described as the ability to add several nucleotides to the 3' end of the synthesized product without the requirement of a template. For example, MMLV RTases preferentially add cytosines to the 3' end of the amplicons (Schmidt & Mueller, 1999). Under specific reaction conditions, $MgCl_2$ and $MnCl_2$ as sources of divalent cations and bovine serum albumin as a stabilizing agent, three to four cytosines can be added by the MMLV RTase to the 3'end of the cDNA.

Another intrinsic property of MMLV RTases is its ability to switch between the templates. This is known as template switching activity and in combination with terminal transferase activity, it can be used to generate specifically tagged cDNA molecules (Zhu *et al.*, 2001). This method requires a special oligonucleotide, which is called a template switching oligonucleotide (TSO). TSO is designed complementary to the overhanging nucleotides, where it serves as a template for elongation of the cDNA product (Figure 5). TSO

primer can be also utilized to add certain desired sequences to both of the cDNA ends, what is often needed in RNA-Seq. Macosko *et al.* (2015) utilize this enzyme's unique property in their Drop-seq sequencing protocol.

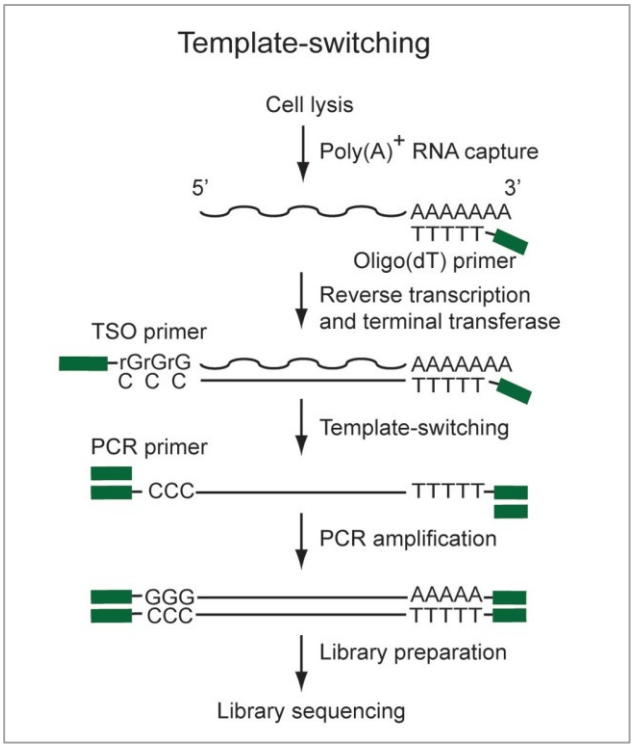

Figure 5: **Template switching mechanism used in constructions of cDNA libraries for RNA-seq.** Reverse transcription is primed with a primer containing PCR handle (green box) and oligo(dT) sequence. Template switching oligonucleotide (TSO) adds PCR handle to the other end of cDNA via complementarity with terminal cytosines (Saliba *et al.*, 2014).

3.4.2 Thermostability and processivity

RNA secondary structures are an obstacle for generation of cDNA. To minimize their impact on the reaction, an elevated reaction temperature is desired since it destabilizes the secondary structure formation. The optimal reaction temperature is considered to be between 50 – 55°C, since it guarantees destabilization of most of the secondary structures, to form primer-template duplexes and RNA is not degraded by the increased temperature (Baranauskas *et al.*, 2012). The enzyme's increased processivity allows for the generation of longer full-length cDNA and shorter reaction protocols (e.g. SuperScript IV, Invitrogen, USA).

MMLV RTase is a suitable candidate for the mutations enhancing its thermostability and processivity, possibly due to its monomeric structure. This might be a reason why most of the commercial enzymes are of MMLV origin (SuperScript III, Invitrogen, USA; Maxima H-, Thermo Scientific, USA; PrimeScript, Takara Bio USA, USA). Multiple point mutations in the study by Baranauskas *et al.* (2012) have been annotated and found to provide superior results when compared to the wild-type enzyme variant. These mutations result in amino acid substitutions and are spread across the enzyme's entire structure. The most enhancing mutations (L139P,

D200N, T330P, L603W) were located in the palm, fingers and RNase H domain and allowed for their combination into one enzyme (the substitutions are encoded in the following manner: native amino acid – position – amino acid substituent). This modified enzyme, when compared to the wild-type variant, had an increased lifetime at 50°C ~ 12-fold (up to 500 min), substrate binding-affinity ~ 50-fold and processivity ~ 65-fold (up to 1500 nucleotides). The mechanism behind its enhanced performance is believed to be achieved by it being more tightly bound to the substrate.

4 Reverse transcription

RT has quickly become a method of molecular biology used across many laboratories all around the world. Most of the primary focus of many gene expression analysis has been on the mRNAs and small RNAs. Throughout time, a myriad of protocols has been developed in pursue to quantify them. As has been noted, several factors may influence the RT efficiency, thus having a further impact on its downstream applications. Both the intrinsic and extrinsic factors of the reaction have a profound impact, creating space for further optimization.

Such occasion is running the reaction at an elevated temperature. Elevated temperature resolves the problems associated with secondary structures but requires the use of a thermostable enzyme. The template's secondary structures have been shown to decrease the RT yield, either by blocking the priming site (Brooks *et al.*, 1995; Kuo *et al.*, 1997) or by slowing down or halting the enzyme (Wu *et al.*, 1996; Suo & Johnson, 1998). Thus, the reaction at an elevated temperature should loosen the secondary structure formation and ease the process of RT. Similarly, mutations increasing other enzyme parameters were introduced, such as processivity (Baranauskas *et al.*, 2012), fidelity (Alvarez *et al.*, 2009) or resistance to inhibitors (SuperScript IV, Invitrogen, USA).

The sample itself is a significant source of variance since the RNA quality restricts the reproducibility and biological relevance of the experiment. It is known that RNA is more prone to changes than DNA, especially in quality. It was described that RNA quality influences the outcome of the study, thus careful sample collection and nucleic acid extraction are of great importance (Vermeulen *et al.*, 2011; Bustin *et al.*, 2009). Additionally, ribonucleases, a type of nuclease catalyzing RNA breakdown, also pose a risk to the RNA integrity. Their negative influence can be prevented by using an RNase inhibitor, ~ 50 kDa protein, that forms a complex with ribonuclease, disabling its function (Dickson *et al.*, 2005).

Furthermore, since RTase requires a primer to initiate the transcription, different priming strategies may be used to produce specific desired effects. Gene-specific primers target specific RNA sequence, in theory reducing the background priming. Oligo(dT) primers aim to amplify mostly polyadenylated transcripts.

Random priming does not target any specific sequence; hence it should prime all sequences equally and deliver the highest overall reaction yield. Combination of priming strategies is possible, depending on the aim of the study.

There are however additional components mixed to the reaction. Ordinary reaction protocol also requires a blend of dNTPs, buffer containing all necessary ions, dithiothreitol, and RNase inhibitor, in some cases the addition of reaction enhancers is also possible. These components may also influence the reaction, e.g. the concentration of Mg^{2+} ions (Goldschmidt *et al.*, 2006). The reaction protocol may also play a role since it is usually RTase specific, some additional steps may be introduced, e.g. annealing of random hexamers at a lower temperature because of their lower melting temperature (SuperScript IV, Invitrogen, USA; AccuScript Hi-Fi, Agilent, USA).

5 Reverse transcription–quantitative polymerase chain reaction

5.1 Polymerase chain reaction

This revolutionary, Nobel prize-winning method of nucleic acid amplification has quickly become essential to molecular biology experiments. Gene expression analysis experiments are no exception, as they also require PCR, where it serves as an important step after RT, amplifying desired cDNA sequences. This method is widely used in the field of gene expression analysis, especially as RT–qPCR, thus basics of this method are shortly summarized in the following paragraphs.

PCR is a method of exponential template amplification, yielding double-stranded DNA product. Requirements for the reaction are: 1) a thermostable polymerase, 2) a blend of dNTPs, 3) a pair of primers flanking the desired sequence and 4) Mg^{2+} ions in the buffer. PCR is performed in multiple cycles, where one cycle consists of denaturation, primer annealing, and primer elongation steps.

Denaturation step should be performed at a temperature high enough to ensure complete dissociation of the double strand. If the temperature is not high enough, the double strand may not fully separate and primers cannot anneal. One needs to be aware of this when amplifying long amplicons (hundreds or thousands bases long). The temperature of primer annealing step should be set slightly lower than the primer's melting temperature. In theory, this ensures the formation of a stable template-primer complex that is recognized by the polymerase. The elongation step is usually performed at 72°C, which is an optimal reaction temperature for *Thermus aquaticus* (*Taq*) polymerase, often used in RT–qPCR. Similarly, elevated temperature helps to melt secondary structures.

5.2 qPCR basics

Quantitative real-time PCR (qPCR) has an additional feature when compared with end-point PCR. The feature is a visualization of amplicon quantity after each PCR cycle. This is achieved by using a fluorescence reporter. In the first cycles, fluorescence signal between amplicon and background cannot be distinguished, however, as the number of amplicons grows, the amplicon signal overcomes the background. With PCR's nature of exponential template amplification, the visualization comes in a shape of exponential curves (Figure 6).

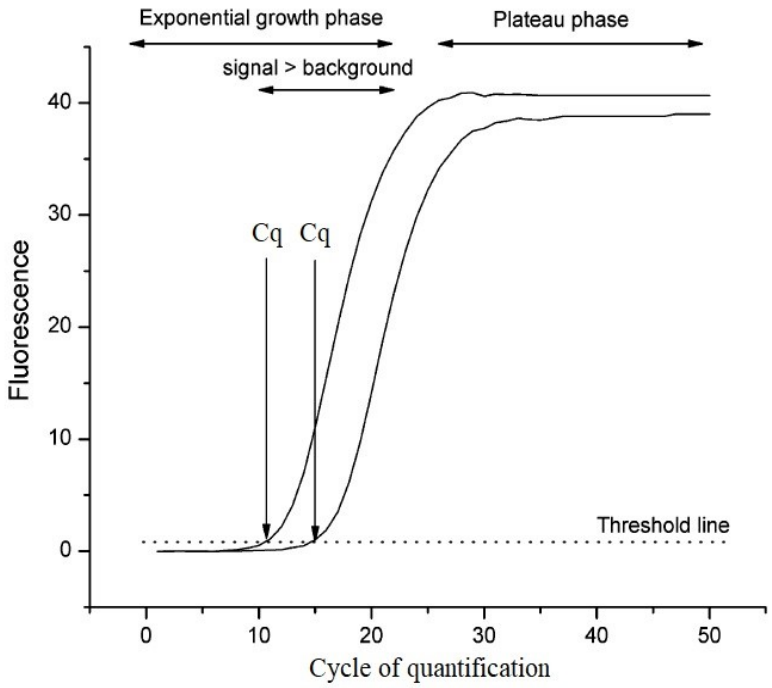

Figure 6: **Real-time PCR amplification curves.** Cycle of quantification (Cq) determines at what point the curve crossed the threshold line. Threshold line must be set above the background noise signal (adapted from Kubista *et al.*, 2006).

The curves enable to distinguish differences in the initial number of template copies, as opposed to end-point PCR, which can only separate positive from negative samples. The differences between the samples are quantified by the cycle of quantification (Cq) values. This value is relative, whereby its informational value is embedded in a comparison with other Cq values read at the same threshold value. Since there are many factors influencing the absolute Cq value, it is not recommended to compare Cqs between the experiments, unless they are normalized for inter-experiment comparisons. Cq represents the number of cycles required to reach the threshold value – a value that can be artificially set. The threshold must be set above the background noise, in the curve's exponential growth phase (Kubista *et al.*, 2006). In practical terms, lower Cq value indicates higher initial template input. Curves reach plateau phase when some necessary reaction component is depleted, such as dNTPs, primers or fluorescent reporter.

The ratio of initial template copies between samples can be expressed as:

$$\frac{[N_0]_B}{[N_0]_A} = 2^{(Cq_B - Cq_A)} \quad (1)$$

where $[N_o]_A$ and $[N_o]_B$ are the initial copy numbers in the samples A and B, respectively and Cq_A and Cq_B are corresponding Cq values. The base of 2 stands for doubling the number of copies per cycle (100 % efficiency). To account for PCR efficiency different from 100 %, the Equation 1 changes accordingly:

$$\frac{[N_0]_B}{[N_0]_A} = (1 + E)^{(Cq_B - Cq_A)} \quad (2)$$

where E stands for PCR efficiency and remaining variables are identical with those from Equation 1. The PCR efficiency can be estimated from a linear function that has been fitted on a series of dilutions. This linear function is called a standard curve (Figure 7). The standard curve's function can be described as:

$$Cq = k \times log_{10}(N_0) + Cq(1) \quad (3)$$

where Cq is the cycle of quantification, k is the coefficient of the variable, $log_{10}(N_o)$ is the number of initial copies in $\log_{10}$ scale and $Cq(1)$ is the intercept, corresponding to Cq of the single initial molecule. The PCR efficiency is calculated using the following equation:

$$E = 10^{-\frac{1}{k}} - 1 \quad (4)$$

where E is the PCR efficiency and k is the coefficient of the variable from Equation 3.

Another measurable outcome, which is often used in studies comparing RTases, is reaction yield. Yield (%) is calculated as a ratio of cDNA to input mRNA molecules. Since prior knowledge of copy numbers is required, comparative studies often use sets of artificial nucleic acids. These artificial nucleic acids (RNA MultiStandard, Roboscreen, Germany; ERCC spike, Thermo Fischer Scientific, USA) have a known number of initial transcripts and known sequences. Reaction yield can be determined as:

$$Yield\ (\%) = \frac{10^{-(\frac{Cq-b}{a})}}{n_{mRNA}} \times 100 \quad (5)$$

where fraction nominator derives from standard curve $Cq = a \times log_{10}(n_{cDNA}) + b$ (Equation 3), where a is the slope, b the intercept, n_{cDNA} the initial number of cDNA molecules and n_{mRNA} the initial number of mRNA molecules.

RT–qPCR is a sensitive and complex method, theoretically any variable may influence the outcome of the reaction. To assure the reproducibility and reliability of the reported results, guidelines ensuring the integrity of the RT–qPCR data were published. The Minimum Information for Publication of Quantitative Real-Time PCR Experiments guidelines (MIQE) is a list of minimum information necessary for publication of qPCR experiments (Bustin *et al.*, 2009).

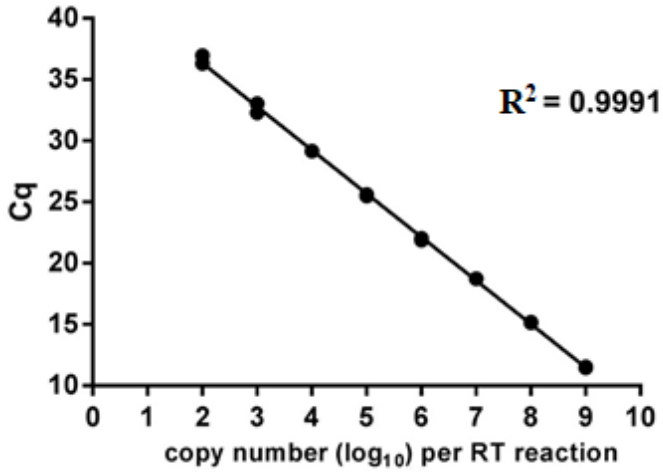

Figure 7: Standard curve plotting copy numbers (in a log_{10} scale) with Cq values. R^2 informs about the goodness of the linear function fit (adapted from Androvic *et al.*, 2017).

5.3 Fluorescence reporters

Use of fluorescence reports enables to visualize the presence of the amplicons. Throughout the years, various fluorescence reporters have been developed, varying in specificity, cost, and complexity of use. Currently, two groups of reporters are used: 1) non-specific labeling chemistries, and 2) specific probes.

5.3.1 Non-specific fluorescence reporters

Non-specific labeling dyes are reporter dyes releasing the fluorescence upon unspecific DNA binding. Currently, mostly asymmetric cyanine dyes are being used, such as SYBR Green. These dyes have near to zero fluorescence when dissolved freely in the solution. However, upon unspecific binding to double strand molecule, presumably to its minor groove (Bengtsson *et al.*, 2003), fluorescence is released. Overall reaction fluorescence signal increases with growing presence of double-stranded molecules. This is also their main disadvantage since they also report the presence of undesired products and primer-dimers. Primer-dimers are a result of unspecific primer binding, the most often with another primer molecule. This, however, does not pose a major problem, since primer-dimers can be revealed by melting temperature analysis. After the last PCR cycle, the temperature is slowly increased, and the fluorescence is measured as a function of temperature. Primer-dimers have a lower melting temperature (Tm) than amplified products, thus their

presence is recognized. In conclusion, the reliability of quantification using non-specific reporter dyes lies in a careful primer design, preventing the primer's self-complementarity.

5.3.2 Specific probes

Specific probes consist of a short nucleic acid sequence with attached fluorescence molecule. The nucleic acid sequence ensures that the signal release is target-specific. However extensive requirements must be met when designing these probes. The molecule must be target specific but at the same time cannot be self-complementary.

TaqMan probes are designed as a linear oligonucleotide, with a fluorophore molecule attached to one end and a quencher to the other end of the sequence. Proximity of the fluorophore and the quencher ensure that no fluorescence signal released. However, when the probe anneals to a template, the fluorophore is cleaved by the polymerase are fluorescence is released (Holland *et al.*, 1991). Molecular beacons use a similar mechanism of quenching the fluorophore, but the spatial proximity is embedded in the beacon's loop structure. Upon annealing to a template, the fluorophore and quencher separate and fluorescence is released (Manganelli *et al.*, 2001). Hybridisation probes consist of two separate molecules. They anneal to a template one following the other. Mechanism of effect lies in excitation of the first fluorophore which transfers the energy to the second fluorophore and the only emission of the second reporter is measured (Caplin *et al.*, 1999).

There is not a single best fluorescence reporting strategy. The choice depends on the subject of the experiment and the cost. Non-specific reporters are cheaper and easier to use but their lack of specificity must be thought of. Problems associated with probes are their challenging design and financial expense, especially in the large-scale studies.

5.4 Inhibitory effects

RT–qPCR is an enzymatically driven reaction, which makes it prone to inhibition from various sources. Some of the known inhibitory components may be already contained in the sample itself. These inhibitors are a broad spectrum of components, varying from heavy metals, highly concentrated metal ions, fats, urea, balt salts, collagen, polysaccharides, phenolic compounds to more complex ones, such as glycogen and humic acids (Rossen *et al.*, 1992*;* Bar *et al.*, 2012; Wilson, 1997; Opel *et al.*, 2010*)*. Similarly, chemicals involved in the sample preparation may have inhibitory effects, e.g. chelators (depleting metal ions necessary for polymerase activity), phenols, KCl, ionic detergents as well as ethanol and isopropanol, substances used in nucleic acid extraction.

All components mentioned above can have an impairing effect on the PCR itself. However since qPCR has an added component of fluorescence, there also might be inhibitors influencing the signal quantification. Unfortunately, this issue remains only partially answered, where some chemicals (hematin, indigo) were observed to limit the dye fluorescence, while others cause Cq shifts and melt curve changes (Opel *et al.*, 2010). To answer this inhibitory effect, use of smaller amplicons may partially relieve the effect but the inhibition should be considered nonetheless. More thorough qPCR inhibition inspection can be achieved by a computational method called kinetics outlier detection (Bar *et al.*, 2012).

Since both RT and qPCR are enzyme-based reactions, it is reasonable to expect that PCR inhibitors may also have an impairing effect on the *in vitro* RT. This may complicate the result evaluation since it is not easily determined at what point was the reaction inhibited. However, the addition of spikes – artificial nucleic acid sequences – may help to resolve this issue (Devonshire *et al.*, 2010).

6 Reverse transcription variance

6.1 Reverse transcriptase reproducibility

Since the reaction is carried out by a single enzyme – RTase, it is reasonable to expect that its performance will impact the reaction's outcome. The first characteristic to consider is the enzyme's reproducibility, as this assumption is crucial to carrying out any RT–qPCR experiment. Enzyme's low reproducibility lowers the credibility of the obtained results, putting into question reliability of any biological variance observed.

The study performed by Bustin *et al.* (2015) on six enzymes, different in origin and characteristics, showed that ΔCq (Cq difference between the lowest and highest replicate) varied between different enzymes, despite the fact that all reactions used an identical template and priming method (performed in 10 RT replicates). In overall, enzyme ΔCqs ranged from 0.4 to 1.74, ReadyScript (Sigma Aldrich, USA) showcasing the smallest differences between RT replicates, while Vilo's (Invitrogen, USA) ΔCq was over four times greater than ReadyScript's. In contrast, ΔCqs of qPCR and pipetting across all six enzymes ranged only from 0.07 to 0.16. These results showcase the fact that RT is responsible for most of the variability in RT–qPCR.

Followed by further analysis carried out on five different mRNA targets, the least (ReadyScript) and the most (Vilo) variable enzymes confirmed their degree of reproducibility with ΔCqs of 0.34 – 1.74 and 0.86 – 3.05, respectively. Since the sample itself can also be a source of variability, validation was performed on another sample. This sample resulted in a lower ΔCq variance of 0.52 – 0.99 and 0.55 – 1.33, respectively. However, it is worth to point out that Vilo consistently recorded lower Cqs – higher yield.

Experiments such as this, conducted in aliquots from a shared template source and diluted to a predefined concentration, allow us to compare RTases directly. Based on their outcome, the use of most reproducible RTase may lower the degree of variance introduced in RT–qPCR experiments. Expressed as a standard deviation (SD) over RT replicates, publications usually report high reproducibility for most RTases tested, often in conditions specific to the study. For example, Ståhlberg *et al.* (2004a) report that RTase reproducibility is gene-related or the work of Levesque – Sergerie *et al.* (2007) presents RTase reproducibility in relevance to concentration of the template or background RNA.

6.1.1 Effect of template concentration

Reproducibility of RT reaction is however not the only assumption one must make. The researcher may also assume that the transcript can be reliably quantified, independently from the template concentration. In other words, the percentage of reverse transcribed RNA remains the same, despite varying template concentration. It has been however shown that the RT yield can vary up to 100-fold (Ståhlberg *et al.*, 2004a), depending on the template concentration and RTase used. In this experiment, RNA MultiStandard (Roboscreen, Germany) was used as a template. RNA MultiStandard is a set of RNA molecules of known concentration and number of copies that can be used to calibrate RT reactions. Even though such high difference has not been reported again, this result calls for a thorough investigation of the phenomenon.

The already mentioned 100-fold difference in reaction yield was reported between 0.4 % yield of AMV (Promega, USA) and 90 % yield of SuperScript III (Invitrogen, USA) (Ståhlberg *et al.*, 2004a). However, as shown in Table 1, RTases did not retain their yields constant across different template concentrations. Mean yield for RNA MultiStandard of SuperScript III outperformed the second enzyme almost 2-fold (83 % for SuperScript III, Invitrogen, USA and 44 % for MMLV, Promega, USA). The lowest recorded mean yield of 2 % was obtained by AMV enzyme (Promega, USA).

Table 1: Absolute RT yields for RNA. Enzymes used are Moloney murine leukemia virus RNase H$^-$ (MMLVH; Promega, USA), OmniScript (Qiagen, Germany), avian myeloblastosis virus (AMV; Promega, USA), MMLV (Promega, USA), Improm-II (Promega, USA), cloned AMV (cAMV; Invitrogen, USA), ThermoScript RNase H$^-$ (Invitrogen, USA), SuperScript III RNase H$^-$ (Invitrogen, USA) (Ståhlberg *et al.*, 2004a).

	Mean (SD) yields[1] (%) at external RNA input (in molecules) of:				Mean (SD)[2] yield for RNA MultiStandard, %
	10^6	10^5	10^4	10^3	
MMLVH	22	50	48	125	40 (16)
Omniscript	7.2	3.1	11.5	66	7.3 (4.2)
AMV	0.4	0.6	4.9	44	2.0 (2.5)
MMLV	32	49	50	110	44 (10)
Improm-II	32	22	12	98	22 (10)
cAMV	6.3	17	35	88	19 (15)
ThermoScript	1.1	9	14	46	8.0 (6.6)
SuperScript III	87	72	90	43	83 (10)
Mean (SD)	24 (29)	28 (26)	33 (29)	78 (32)	28 (27)

[1] *RT yields of RNA prepared from liver and spleen. Note the markedly higher yields at an input of 10^3 RNA molecules.*

[2] *RT yield for samples containing 10^4-10^6 RNA MultiStandard molecules.*

In a different study, conducted by Levesque-Sergerie et al. (2007), SuperScript II (Invitrogen, USA) significantly outperformed other enzymes for low mRNA input (1 fg *EGFP* mRNA = 2.6×10^3 transcript copies), with yields ranging from 19.8 % to 102.4 %. Its outstanding performance was detected for a wide range of background RNA (Figure 8). However, it is important to note that investigation of SuperScript II standard curves suggests an artificially enhanced results, possibly due to incorrectly performed standard curve preparation (Miranda & Steward, 2017). In conclusion, this observation should be taken with caution. For high mRNA inputs (1 pg *EGFP* mRNA = 2.6×10^6 transcript copies), SensiScript (Qiagen, Germany) and PowerScript (Clontech, USA) enzymes can be declared as having the best-performing RT yields with outputs up to 50.41 % and 59.43 %, respectively (Figure 8). Since these measurements were primarily conducted on an exogenous target, additional validation on an endogenous sequence was performed. Detection of endogenous *GNDPA* mRNA, present in RNA background, confirmed SensiScript's performance with 42.49 % yield. In the context of their experiment, a number of *GNPDA* copies can be considered as a middle-abundant target.

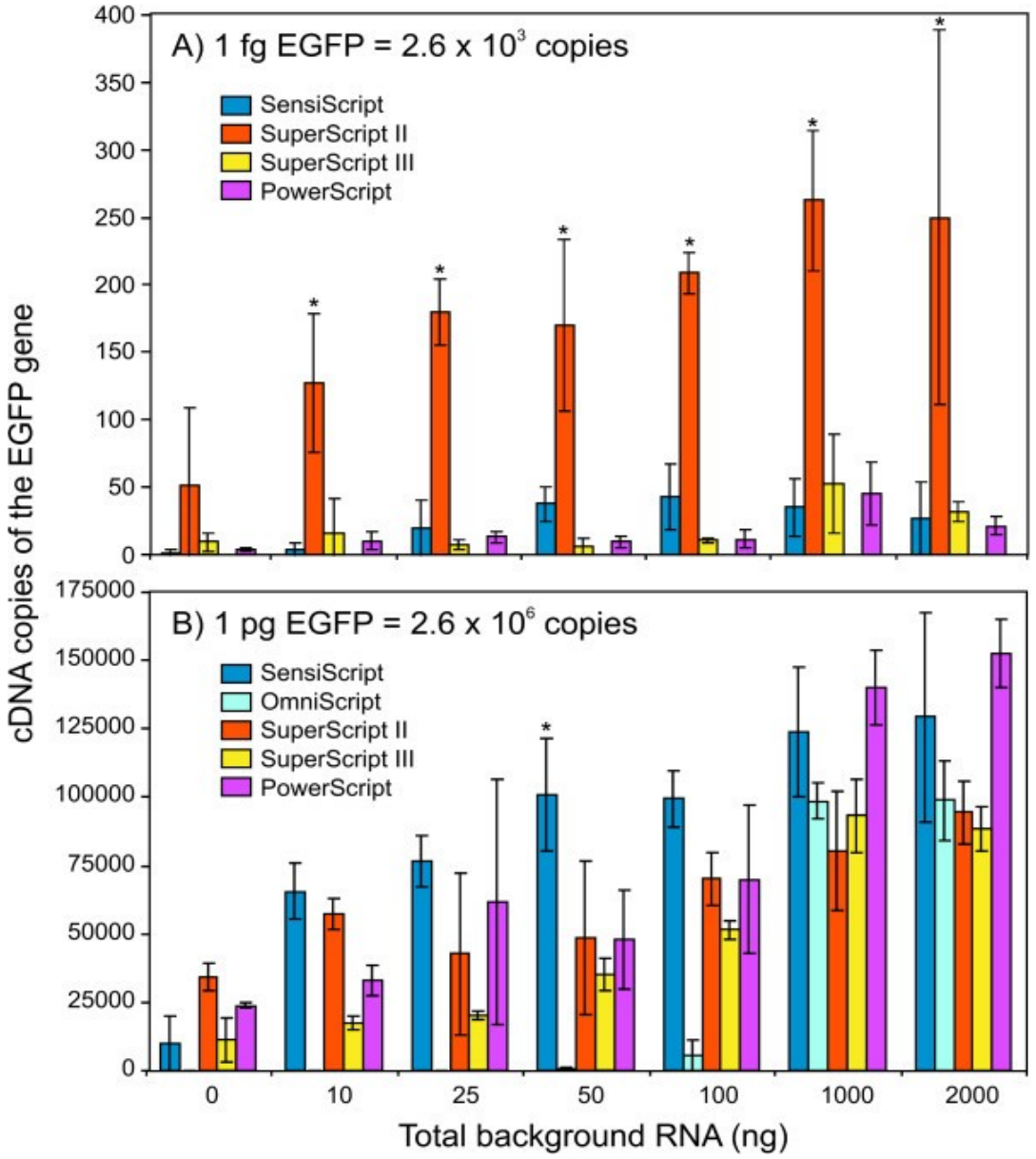

Figure 8: Quantitative measurements of RT reactions performed with five commercial systems spiked with low-abundant (A) and high-abundant (B) *EGFP* mRNA template. qPCR measurements were conducted using 1/10 of RT reaction. Absolute qPCR copy numbers are reported. cDNA produced by OmniScript, in low transcript abundant reaction (A), was undetected. Reactions were performed in triplicates. Absolute values of *EGFP* copy numbers were determined from a standard curve of purified *EGFP* DNA fragment. Significantly different results ($p < 0.05$) are marked with an asterisk (*) (Levesque-Sergerie *et al.*, 2007).

In order to verify the superior yield of SuperScript II in low-abundant templates, a repeated comparison was performed. Performance of SuperScript II (Invitrogen, USA) and SuperScript III (Invitrogen, USA) was tested at low (6.1×10^2 copies) and high (6.1×10^5 copies) template concentrations (Miranda & Steward, 2017). On contrary, this experiment showed that a significant difference was observed only at a higher concentration (*t-test,* $P < 0.001$). In three separate experiments, SuperScript II recorded a yield mean varying from 35 % to 69 % across a wide range of initial template copies. This mean variation was correlated to template concentration ($r = 0.52$, $P = 0.003$), however, this should be regarded with caution, because this

correlation may have been strongly influenced just by one of the experiments (Experiment 1) (Figure 9). In the other two experiments, the SuperScript II did not report significant correlation between efficiency increase and higher template input (Experiment 1: r = 0.95, $P < 0.001$; Experiment 2: r = 0.59, $P = 0.07$; Experiment 3: r = -0.03, $P = 0.95$). These findings raise concerns about the choice of specific RTase for practical applications, especially when additional factors, such as targets abundance or amount of RNA in the reaction, may impact the results in an unpredictable manner.

Both studies of Ståhlberg *et al.* (2004a) and Levesque-Sergerie *et al.* (2007) shared multiple reaction parameters, e.g. use of artificial RNA input, template concentration range (10^3 – 10^6 copies), RNA background (Ståhlberg's 43 ng/µl and Levesque-Sergerie's 50 ng/µl); but also varied in some, e.g. priming strategy (random hexamers and oligo$(dT)_{12-18}$, respectively), primer concentration (5 µM and 10 µM, respectively). Additionally, both studies tested OmniScript enzyme (Qiagen, Germany). To a general surprise, however, OmniScript's yields for 10^3 input molecules were 66 % and 0 % for Ståhlberg and Levesque-Sergerie, respectively. Whereas for 10^6 input molecules it was 7.2 % and 38.46 %, respectively. Discrepancies on such scale, even when some non-identical reaction conditions are met, demand not only further investigation of factors influencing RTase efficiency but may also report low reproducibility of the results between laboratories.

6.1.2 Variability caused by background RNA

In the study of Levesque-Sergerie *et al.* (2007), reaction yields closely followed growing presence of RNA background (Figure 8), where 50 ng/µl (1000 ng in 20-µl reaction) was reported to be a sufficient background RNA concentration to maximize reaction yield. This finding was once more confirmed in a similar range, reflecting inhibition by background RNA past 50 ng/µl concentration (Miranda & Steward, 2017).

The positive effect of adding background RNA was not reported to be equal for all types of RNA. Levesque-Sergerie *et al.* (2007) state that when total RNA was substituted for tRNA, the improved performance was not observed. Despite this, tRNA was previously reported as a component ensuring linear sample dilution (Ståhlberg *et al.*, 2004b). In light of these findings, it can be stated that the molecular role of nucleic acid background in RT is only yet to be understood.

6.1.3 Bayesian modeling

A different approach for measuring RTase-to-RTase variance was developed using Bayesian modeling, where the main goal was to minimize bias introduced by target genes, sample size and most importantly, the laboratory carrying out the experiment. In return, this model should evaluate enzyme- and gene-specific variances for infinite population models (Lindén *et al.*, 2012).

The model was built on three levels, where enzymes composed the highest level of the hierarchy. The middle level was composed of five RT replicates by each enzyme, what was represented as parameter E. This parameter E describes the enzyme's performance, such as efficiency and reproducibility. The lowest level was four studied housekeeping genes present in total RNA. Testing eight RTases of various origin and RNase H activity came to a conclusion of moderate enzyme efficiency variance across enzymes studied, with exception of one, which is no longer available on the market (StrataScript RT) (Figure 10).

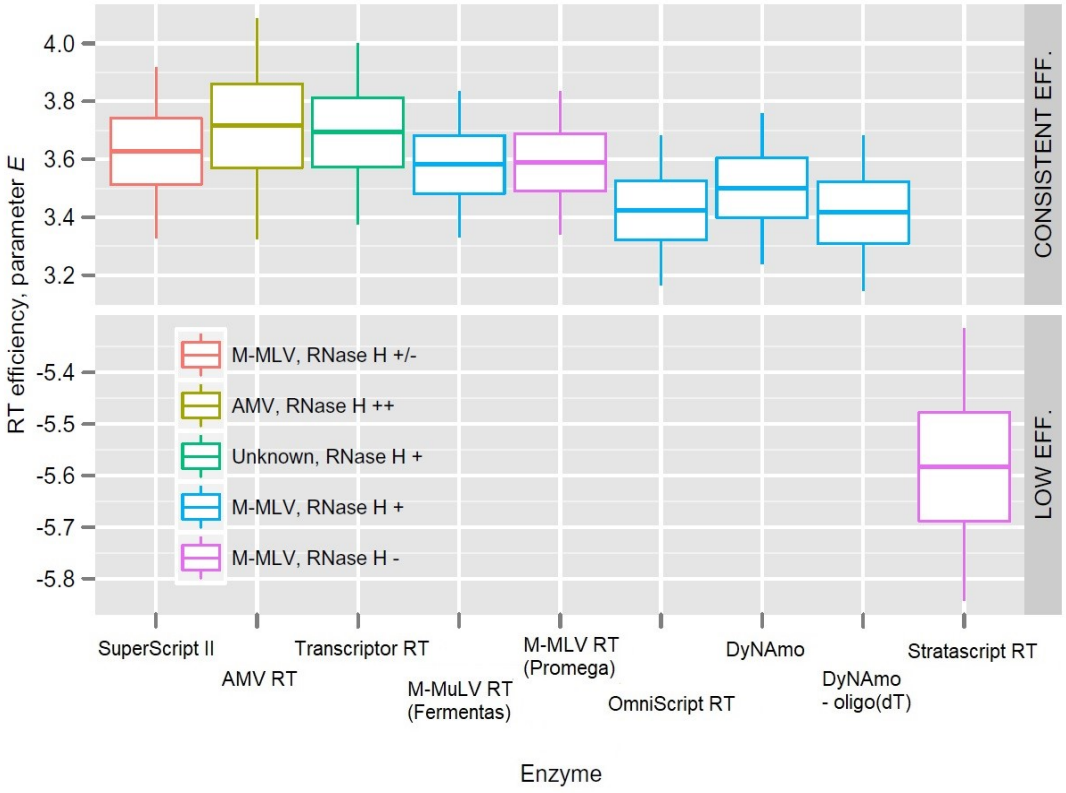

Figure 10: **Parameter E interpreted as RT efficiency of different RTases.** Middle line denotes density median, upper and lower box hinges 0.75 and 0.25 quantiles, respectively. Whiskers span from 0.05 to 0.95 quantiles. The legend denotes enzymes origin and RNase H activity (++ = strong, + = present, +/- = reduced and - = absent). Enzymes are as follow: 1. SuperScript II, Thermo Fischer, USA; 2. AMV RT, Finnzymes (Thermo Fischer Scientific), USA; 3. Transcriptor RT, Roche, Switzerland; 4. M-MuLV RT, Fermentas (Thermo Fischer Scientific), USA; 5. M-MLV RT RNase H^-, Promega, USA; 6. Omniscript, Qiagen, Germany; 7. DyNAmo, Finnzymes (Thermo Fischer Scientific), USA; 8. DyNAmo, Finnzymes primed only with hexanucleotides (Thermo Fischer Scientific), USA;9. Stratascript RT, Stratagene (Agilent), USA (adapted from Lindén *et al.*, 2012).

Regarding general enzyme-specific reproducibility, expressed as median of the posterior probability density of infinite populations SD parameter (σ_e), enzymes could be divided into two categories ranging 0.13 - 0.20 (enzymes 4 to 8) and 0.29 – 0.52 (enzymes 1 to 3), respectively (Figure 11). In practical terms, this parameter informs about the expected enzyme-specific SD, regardless of the template. Although the differences were not significant, σ_e of the least varying enzyme (M-MLV RT) was smaller than σ_e of the most

varying enzyme (AMV RT) with a probability of 0.92 and smaller than σ_e of the second most varying enzyme (Transcriptor RT) with a probability of 0.78.

In conclusion, these findings report the existing differences between the enzyme-specific reproducibility, independent of other influencing factors. However, there was no statistically significant difference between the enzymes measured. In addition, as shown in Figure 11, RNase H only has a little effect on the enzyme's performance.

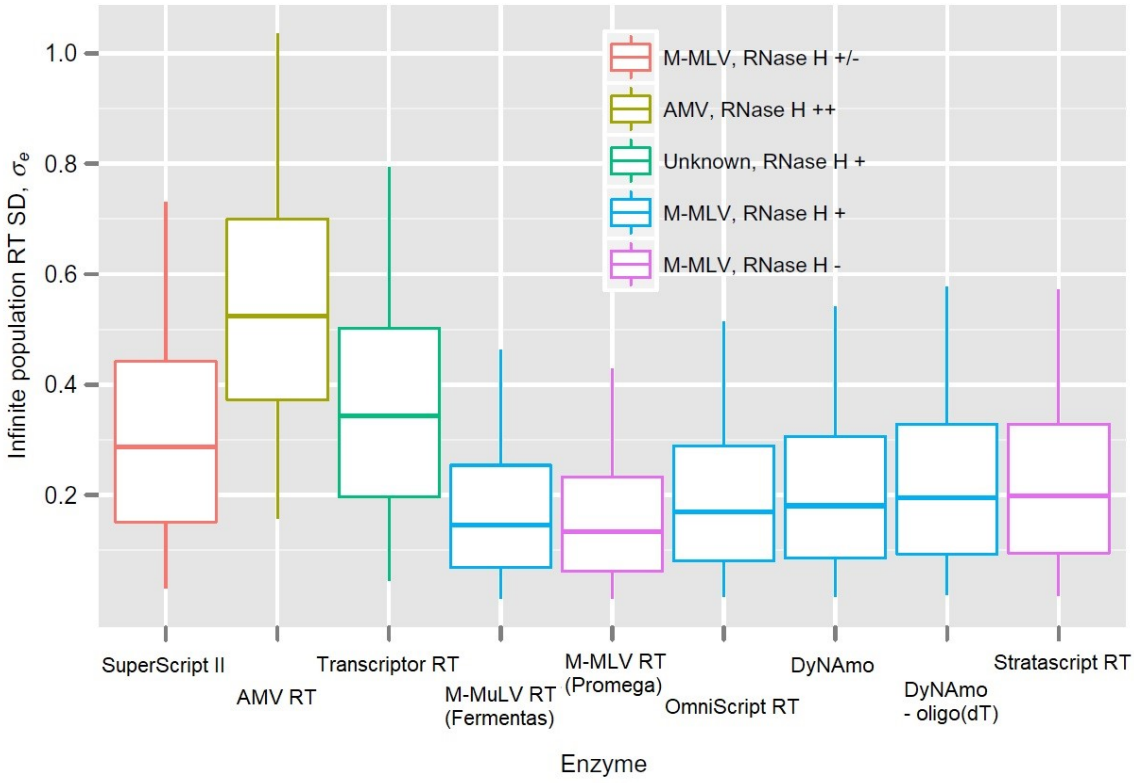

Figure 11: Posterior probability density estimates of the enzyme-specific infinite population RT SD parameters (σ_e). Results were obtained on four control genes. The middle line is density median, upper and lower box edges represent 0.75 and 0.25 quantiles, respectively. See Figure 10 for legend and enzyme description (adapted from Lindén *et al.*, 2012).

6.2 Gene- and sample-specific variability

Even though Lindén *et al.* (2012) report the limited degree of variance introduced by the enzyme, their model revealed that gene-specific factors were main contributors to the variances in the infinite population model. Meanwhile, only modest differences were seen between finite population variance of gene-specific factors. The median finite population gene-specific parameters expressed as coefficient of variation (CV) were 25 % for *Gapdh* and *Pgk1*, 55 % for *Actb* and 70 % for *Sdha*. For infinite population, it was 9 % for *Gapdh*, 11 % for *Pgk1* and very large, uninformative CVs for *Actb* and *Sdha*. The model, in general, informs about good RT-to-

RT reproducibility for stable transcripts, but at the same time predicts unsure results for *Actb* and *Sdha* genes. In conclusion, this model helped to point out the lack of the stability of some possible reference genes (*Actb*), which may become even more pronounced with use of infinite population model.

Concerns about gene-specific variability were raised also earlier by Ståhlberg *et al.* (2004b), when they showed that gene related estimation of cDNA copy number for most of their tested genes (*β-tubulin, CaV1D, GAPDH, Insulin II*) reported variance under 8 %. However, the least expressed gene in their study (*Glut2*) reported an estimate variation as high as 26 %.

In accordance with this trend, Ståhlberg *et al.* (2004a) also notified the gene-specific RT efficiency variance. Across eight enzymes tested, three genes were expressing little variance (*HTR1a*, *HTR1b,* and *HTR2b*), one gene moderate variance (*GAPDH*) but for two genes (*β-actin*, *HTR2a*) up to 91-fold yield difference was observed between two enzymes (SuperScript III, AMV). In perspective, those same enzymes yielded only 1.14-fold difference for *HTR2b*.

Reliable template quantification can be also interpreted in relative expression profiles. Relative expression profile is calculated by dividing copy numbers of the more expressed gene by the less expressed one. Bustin *et al.* (2015) compared relative expression profiles of two gene pairs, *CDK2/RBL1* and *MAX/MYC* across four different high-quality RNA samples. Significant gene expression correlation was remarked in both gene pairs, with exception of one sample in the *MAX/MYC* pair. However, only *CDK2/RBL1* pair had similar relative expression levels and differences in fold change (Δ fold change) across all RNA samples. In the case of the *MAX/MYC* pair, the variability was significant (average relative expression level of 0.8 with SD of 0.5; average Δ fold change of 3.5 with SD of 3.0), as can be seen in Figure 12. The inconsistency of *MAX* target was confirmed on repeated measurement of 35 RT replicates in the *CDH1/MAX* pair. *CDH1/MAX* pair however retained good correlation coefficient. In practice, it is advised to regard reported inconsistency with caution, since data investigation showed that as few as three out of 40 replicates were notably responsible for inconsistency in *MAX/MYC* measurements.

Throughout various experiments, it has been noted that mRNA targets are not evenly reverse transcribed. Although some authors suggest that RT reaction simply favors transcription of highly-abundant over less expressed targets in the presence of low background RNA (Levesque-Sergerie *et al.*, 2007; Ståhlberg *et al.*, 2004b). This was not uniformly the case in the previous study (Ståhlberg *et al.*, 2004a), since not all low-expressed targets reported equal degree of variation. There is no clear answer what causes this phenomenon. The practical validation of reproducibility is therefore of prime importance for any target that will play a major role in data analysis, e.g. reference gene.

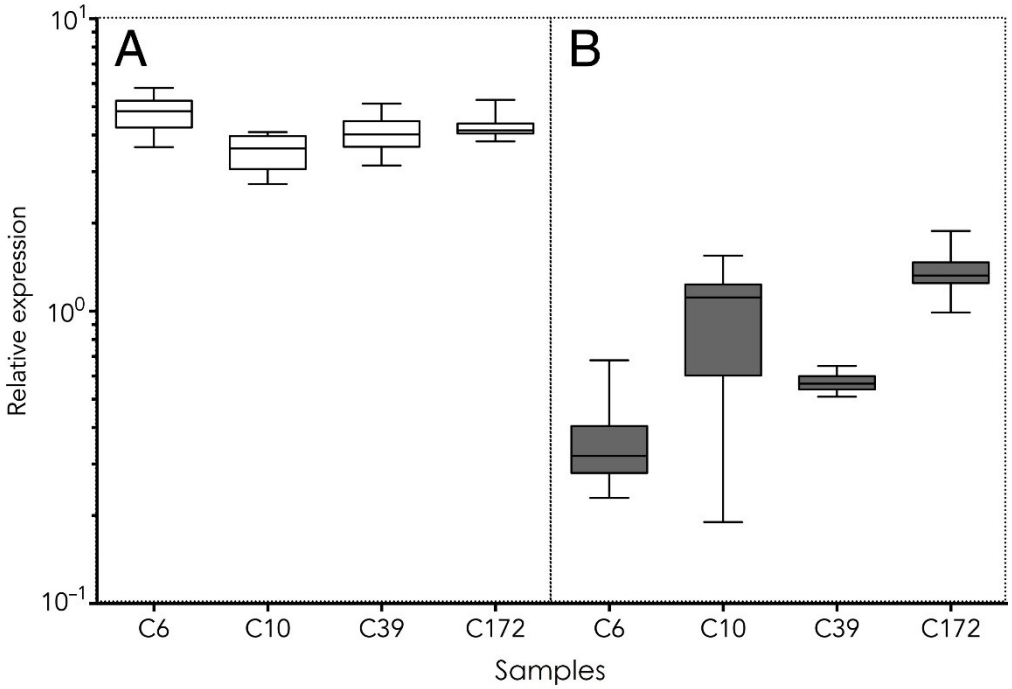

Figure 12: Expression levels of (A) CDK2 relative to RBL1 and (B) MAX relative to MYC carried out on 10 replicates per sample (C6, C10, C39 and C172). Relative expression calculation was performed by dividing the copy numbers for CDK2 and MAX by the copy numbers of RBL1 or MYC, respectively. Δ fold change for each sample was calculated by division of the highest relative expression with the lowest relative expression, for each sample separately. The middle line denotes the median value, upper and lower box edges 0.75 and 0.25 quantiles. Minimum and maximum expression levels are represented by error bars (Bustin *et al.*, 2015).

6.3 Priming

Choice of RT primers is one of many options one can make when preparing the reaction. However, substantial discrepancies between priming strategies have been reported, raising awareness of consequences a researcher's decision may lead to. Despite this, only a few experiments were dedicated to enlarging our understanding of this mechanism.

6.3.1 Random primers

The least specific priming method has often been the point of discussion. In theory, oligonucleotide synthesized by random nucleotides should have a potential to prime all RNA transcripts equally and deliver a higher yield than other priming methods. However, there is not an ultimately best priming method to be used in all cases and random primers are not an exception to this. For example, rRNA, being the most abundant RNA type in the cell, can and will be primed by random primers even though it usually is not a subject of the experiment. Additionally, there is not a theoretical boundary preventing single RNA molecule to be primed

from multiple sites, eventually leading to generation of multiple cDNA copies from one RNA transcript, thus falsely overestimating its presence in the sample.

In the study performed by Zhang & Byrne (1999) the calculation of initial mRNA copy numbers varied significantly, depending on the primers used. Copy numbers obtained with random hexamers reported 5- and 19- fold difference when compared with specific hexamers (primers targeting specific sequence, six nucleotides long) and specific 22-mers (primers targeting specific sequence, 22 nucleotides long), respectively. The calculation method utilized ratio of target and standard cDNA molecules interpreted as optical density on an electrophoresis gel. However, weakness of this interpretation is embedded in the design of standard's priming sites, since the sequence inspection revealed that 3' end of the standard molecule was too short. 3' end was designed to be PCR priming site but because there was not downstream sequence after it, random primers failed to reverse transcribe the PCR priming site reliably. This resulted in underperformance of standard molecules and false overestimation of random primer yields.

Ståhlberg *et al.* (2004b) shed more light on the efficiency of priming strategies, comparing random hexamers, oligo(dT), gene-specific primers and a mixture of gene-specific primers (GSP), where no priming strategy outperformed the others for all genes measured (Table 2). Random hexamers were the optimal alternative for three genes out of five, but significantly only for the case of *CaV1D*.

Table 2: Reliance of RT reaction on priming strategy used. For each gene, the lowest Cq value is underlined. The priming strategy for β-Tubulin, CaV1D and Insulin II are better than its counterparts with 99 % confidence. The maximum ΔCq was calculated as difference between the best and worst Cq for each gene (Ståhlberg *et al.*, 2004b).

Priming strategy	Cq				
	β-Tubulin	*CaV1D*	*GAPDH*	*Insulin II*	*Glut2*
Random hexamers	19.5	26.5	15.8	16.9	27.5
Oligo(dT)	18.1	28.8	16.6	15.9	28.4
Gene-specific primers	18.8	28.7	16.4	17.4	31.8
Mixture of 5 GSP	19.1	27.9	16.2	16.6	29.3
Maximum ΔCq	1.4	2.3	0.8	1.5	4.4

Random primers, however, do not have to be only six nucleotides long. The effect of varying length of random primers was studied by Stangegaard *et al.* (2006), on primers ranging from 6 to 21 nucleotides long. As a result, they observed that random pentadecamers yielded two times more cDNA copies than random hexamers. This 2-fold yield increase was also confirmed across different RNA templates or with three different RTases (SuperScript II, AMV, MMLV). The possible explanation for this may be the formation of stronger duplexes between template and primer since pentadecamer's melting temperature (approximately 40°C) is

closer to reaction temperature (42°C) than random hexamers' (approximately 20°C). Another possible explanation for this is that duplex may be easier to recognize by the enzyme, thus priming the reaction more efficiently.

6.3.2 Oligo(dT)

Priming method of enhanced specificity, when compared to random primers, amplifies transcripts from 3' end poly(A)-tail. In theory, oligo(dT)s should capture all polyadenylated mRNAs, leaving rRNA and other non-polyadenylated RNAs unprimed (e.g. viral, prokaryotic or histone-specifying). Poly(A)-tail priming requires starting material of very high quality since poly(A)-tail is prone to fragmentation and degradation. This may limit the sample preparation methods, for example for formalin-fixed paraffin-embedded (FFPE) samples oligo(dT) priming is not a suitable strategy (Zeka *et al.*, 2016).

Despite all the requirements, oligo(dT) may exhibit efficient performance across wide range of input material and different transcripts. Lekanne Deprez *et al.* (2002) reported almost linear amplification for each of five mRNA targets with a correlation coefficient (*r*) of at least 0.99 across input range from 0.125 to 4 µg of total RNA. However, it is important to note that in this experiment the oligonucleotide did not consist of thymine nucleotides only, but primer's 3' end was extended by an anchor sequence – VN nucleotides (V – not thymine, N – random nucleotide). In theory, anchor sequence should secure annealing precisely on the border between transcript's sequence and poly(A)-tail.

Ståhlberg *et al.* (2004b) results confirmed oligo(dT)'s good performance, reporting oligo(dT) as optimal priming strategy for two genes out of five, both with 99 % confidence (Table 2).

6.3.3 Gene-specific primers

To reverse transcribe a specific sequence, the transcript can be primed with a pre-designed sequence-specific oligonucleotide. The downside of this strategy is, however, the cost (each sequence requires a unique primer), specificity validation and possibly inter-primer cross-reactivity when a mix of gene-specific primers is used in one reaction.

mRNA copy numbers reported by Zhang & Byrne (1999) showed reaction's varying yield and accuracy, dependent on the specific primer's length and elongation temperature. Longer specific primers led to the synthesis of full-length cDNA molecules what helped to increase reproducibility and precision of RT. Additionally, the authors recommend using longer primers when the reaction is performed at elevated temperatures. Their findings suggest that short primers produce a large portion of truncated cDNA molecules, whereas longer primers (with higher Tm) produce more full-length molecules, introducing less gene expression profile alterations.

Nonetheless, the effectiveness of gene-specific priming varies from study to study. Lekanne Deprez *et al.* (2002) reports that gene-specific priming outperforms oligo(dT) and random priming in terms of yield, whereas the study by Ståhlberg *et al.* (2004b), single gene-specific primers or a mix of five primers reported higher Cqs for all 5 genes measured (Table 2).

A possible explanation for this disagreement may be different primer concentration used. Whereas Lekanne Deprez *et al.* (2002) primed the reactions with a concentration of 4 and 12.5 μM, Ståhlberg *et al.* (2004b) used 1 and 50 μM for gene-specific primers and random hexamers, respectively. This may answer discrepancies observed since Miranda & Steward (2017) report that increase in primer concentration resulted in lower Cqs for both priming methods. Furthermore, both strategies responded similarly to the concentration increase - linear regression slopes were indifferent (*t*-test, n = 12, $P = 0.21$). Both strategies saturated with similar Cq values, but random priming required up to 5-fold higher concentration to obtain the same Cq. Similar Cq at saturated primer concentration was also confirmed using wide RNA concentration range (5×10^2 - 5×10^6 copies per RT reaction) for both priming strategies.

7 High-throughput gene expression analysis

Large gene-scale qPCR studies are impractical for reasons of necessary labor time, cost and possible bias introduced by the researcher, either when choosing the set of genes or manually performing the measurements. With transcriptome-wide sequencing, thousands of genes can be studied simultaneously, what enables to study gene interactions on a large scale. Decreasing sequencing costs make it a more accessible platform for many research teams, eventually accelerating its further development and therefore the amount of knowledge obtained with it.

RNA-Seq is especially useful in studies of single-cell transcriptomes, as it allows to describe heterogeneity between individual cells in heterogeneous cell populations (Patel *et al.*, 2014; Zeisel *et al.*, 2015). Multiple protocols have been developed, varying in their accuracy and sensitivity (Figure 13) (Svensson *et al.*, 2017). Single-cell RNA-Seq (scRNA-Seq) protocols also find use in research of rare cell types, such as circulating tumor cells or stem cells (Yu *et al.*, 2012; Yan *et al.*, 2013).

Since conventional sequencing requires DNA as a template, RT is a necessary step in the sample preparation. In order to simplify and strengthen the RNA-Seq data evaluation, modifications to conventional RT protocol were introduced, especially focused on primer sequence used. Such modifications are the addition of T7 promoter, cell barcode, a unique molecular identifier (UMI), PCR handles and eventually any other desired sequence in the sequence of new synthetized cDNA. The information about transcript's cell of origin

is stored in cell barcode, UMIs are used to remove bias introduced by PCR amplification of cDNA molecules and PCR handles serve for specific amplification and sequencing (Saliba *et al.*, 2014).

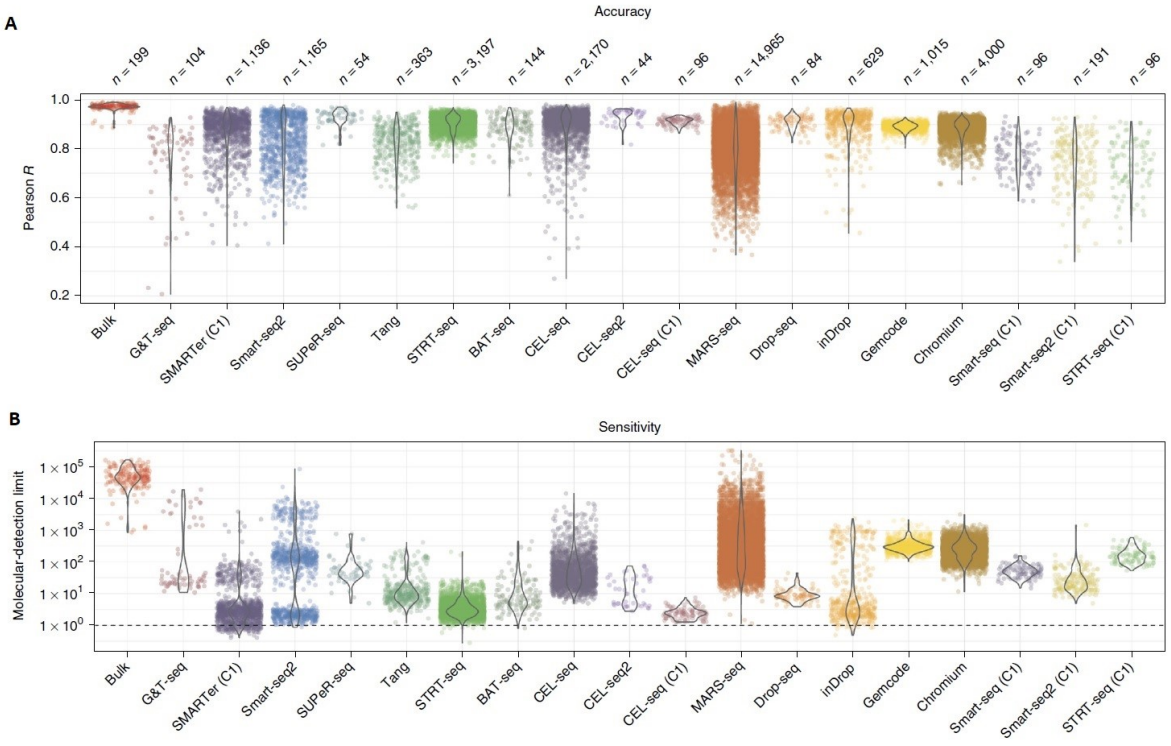

Figure 13: **Performance comparison of scRNA-Seq protocols.** (A) Accuracy based on Pearson correlations (R), calculated from number of observed and input ERCC spike molecules. (B) Sensitivity of the method validated as spike input level with 50 % probability of ERCC spike detection. n stands for number of samples (Svensson *et al.*, 2017).

Since RNA-Seq is a continually improving method and there is not an exclusively best RT method for RNA-seq, several methods of cDNA generation are available. A method of PCR amplification shares many similarities with ordinary RT. It relies on the addition of PCR handles onto the 5' and 3' cDNA ends, allowing for later PCR amplification (Figure 5). RT is initiated by priming with oligonucleotide consisting of oligo(dT) and desired sequence (e.g. PCR handle). Afterwards the primer is elongated, and cDNA is produced with the extension of few dCTPs, which are added by RTase's intrinsic property of terminal transferase activity. This extension can serve as a template for TSO primer. Since TSO primer extends the template for RTase, MMLV's template switching mechanism takes place and PCR handle is reverse transcribed into cDNA 3' end. This step is followed by PCR amplification and library sequencing (Zhu *et al.*, 2001). Drop-seq protocol developed by Macosko *et al.* (2015) is utilizing this method in practice.

There are also protocols avoiding the need for template amplification and instead of it they use the process of *in vitro* transcription, e.g. Cel-Seq2 and MARS-Seq (Hashimshony *et al.*, 2016; Jaitin *et al.*, 2014). The primer used to initiate the first RT step consists of multiple sequences as outlined in Figure 14. T7 promoter is a necessary sequence for *in vitro* transcription by T7 polymerase; Illumina 5' adaptor is sequence specific to sequencing technology; UMI's ensure that each reverse transcribed mRNA is counted precisely once; barcode is sequence unique to each cell that is being sequenced, allowing to assign each transcript to its host cell; oligo(dT) primes the RT of mRNA. The protocol follows the steps described in Figure 14. In step 5, 3' overhang on random primers is sequence necessary for the sequencer.

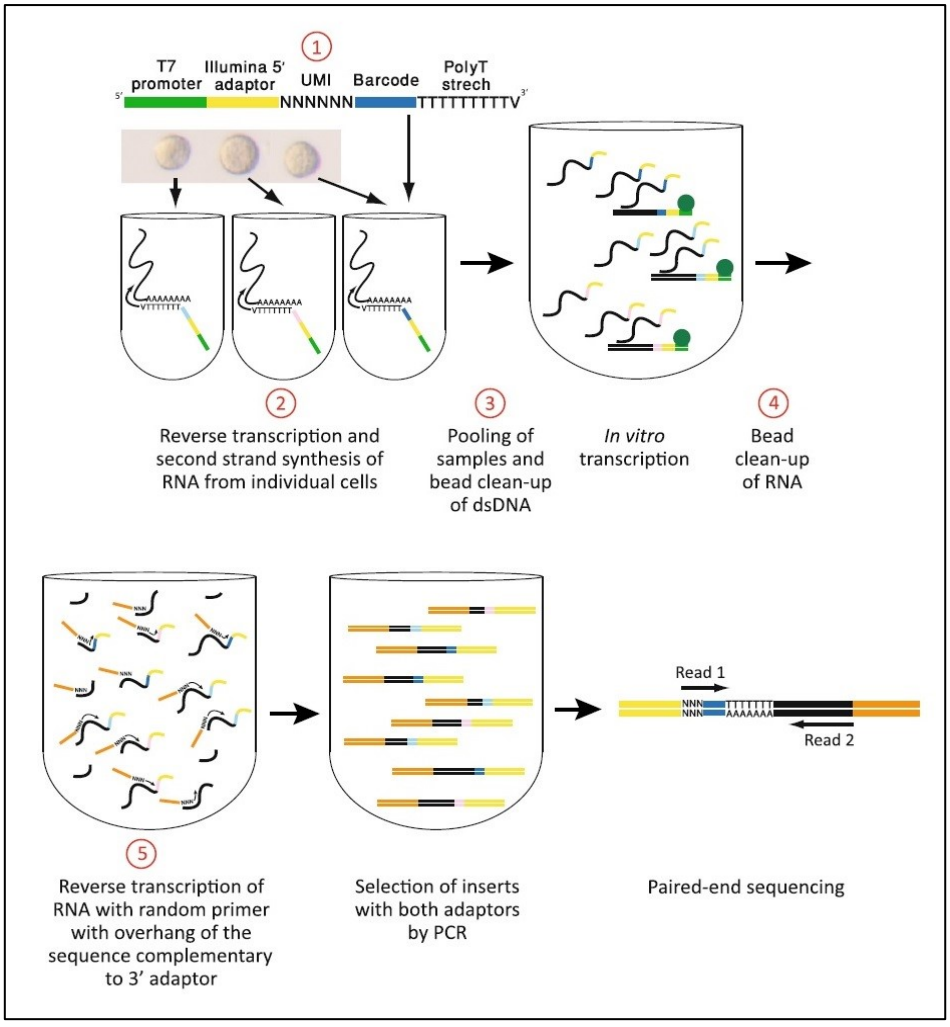

Figure 14: A simplified representation of CEL-Seq2 protocol (Hashimshony *et al.*, 2016).

In cases where the non-polyadenylated RNAs are a subject of the study, such as miRNAs (microRNA) and snoRNAs (small nucleolar RNA), a method of small-RNA transcriptome sequencing was developed, introducing additional ligation and enzymatic removal steps (Faridani *et al.*, 2016). The steps of the protocol are outlined in Figure 15. The protocol utilizes ligation of adaptors to all RNAs containing 5' phosphate and 3' hydroxyl group and small RNAs are selected afterward computationally. Since ligation used in the protocol is

specific to single-stranded RNA, adaptor ligation to abundant 5.8S rRNA can be prevented by using specific oligonucleotides. RT reaction is then initiated from the sequence ligated on 3' end of small RNAs. Indexed primer serves the same purpose as barcode sequence; allows for the pooling of the samples since the cell of origin can be assigned computationally.

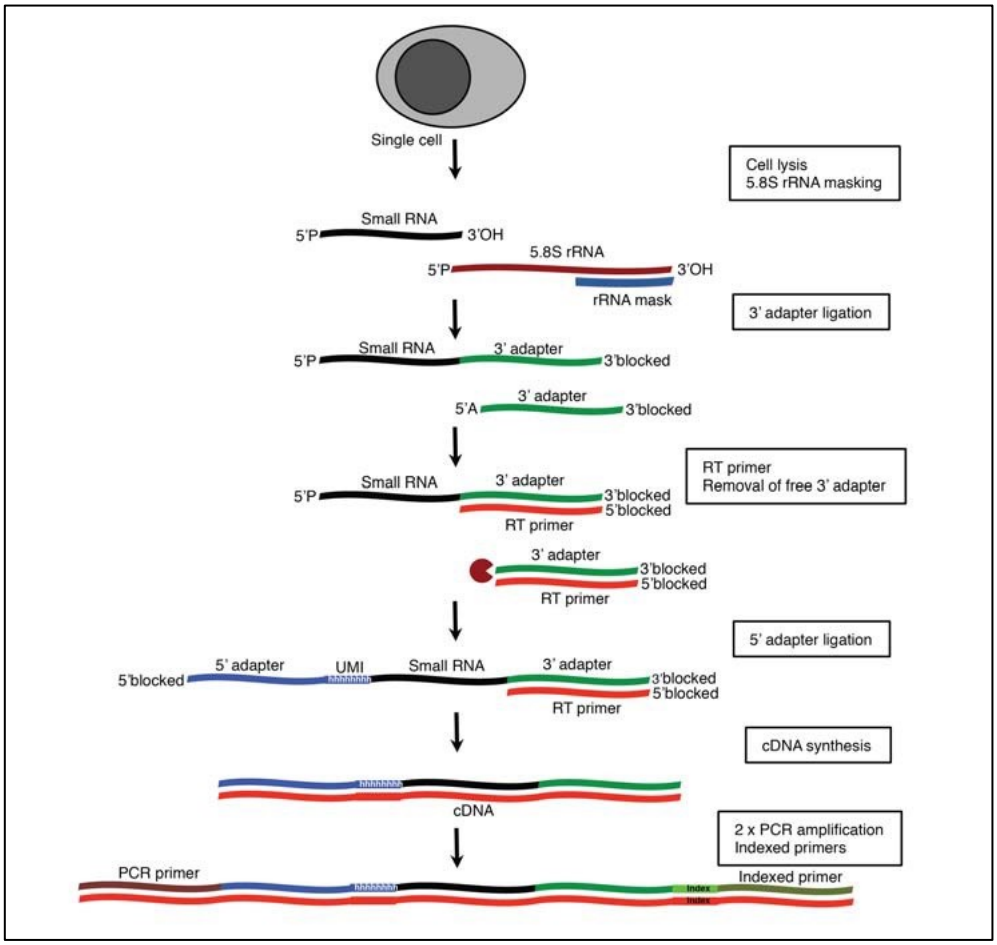

Figure 15: **An outline of small-RNA-Seq protocol.** The majority of rRNA is masked with a masking primer, blocking the ligation of 3' adapter. The UMI is contained in the 5' adapter (Faridani *et al.*, 2016).

RNA-Seq's high precision in combination with transcript labeling allows for precise data evaluation and large-scale study at the same time. Despite this, RT is still a necessary step and as it has been summarized in this thesis, RT can account for significant differences and introduce bias. In pursuit of overcoming these drawbacks, several direct RNA sequencing methods were developed (Ozsolak *et al.*, 2009; Hickman *et al.*, 2013).

8 Conclusions

The aim of this thesis was to summarize the role of RTase in gene expression analysis. RTase is an enzyme carrying out a unique reaction, synthesizing DNA from RNA template, which can then be used for transcriptomic analysis. Since RNA cannot be quantified by the methods of qPCR or sequencing, RT mainly serves the purpose of generating the necessary template. This inevitable step is less precise than its downstream applications and introduces bias.

RTase, as a main component of the reaction, is a partial source of this variation. RTases of MMLV origin, in general, prove to perform better than its AMV counterparts. This can be possibly attributed to their structure since it is easier to introduce enhancing point mutations into the structure of monomeric MMLV enzymes without disrupting its functionality. Additionally, mutations in RTase's RNase H activity were also shown not to deliver a significant impact.

The main source of RT variance is attributed to templates of the experiment – samples and transcripts. The main reason is that not all transcripts are being reverse transcribed with equal efficiency. It is not yet fully understood what lies behind this phenomenon, but the presence of secondary structures may partly explain this problem. The complete mechanism, however, remains to be discovered. Additionally, RT reaction seems to be less sensitive to low expressed transcripts, transcribing them with lower efficiency than the highly expressed ones. This issue can be minimized by use of background RNA and suitable RTase.

No priming strategy is proved to be superior to the others, thus the final choice of primers should be relevant to the goal, cost, and design of the study. Although some experiments require specific priming method, partial adjustments can be performed, such as the use of random pentadecamers instead of random hexamers or addition of anchor sequence to the oligo(dT). Random primers produce the highest yield, are the cheapest option, but lack the specificity. Oligo(dT)s theoretically transcribe only polyadenylated RNA and can deliver good yields, but RNA secondary structures and material quality may limit the efficiency of their use. The most specific product can be synthesized with gene-specific primers. However, the primer design, validation, and cost may become an obstacle to their use.

In practice, the best performing RT is specific to experimental design. Empirical evaluation can be conducted on a pilot study of few enzymes prior to the main experiment. This study can help to address multiple issues that may arise later: 1) addition of RNA spike to the template may determine RT efficiency at different template concentrations, 2) review the impact of background RNA on the reaction outcome, 3) optimize priming strategy (both in terms of yield and purposes of downstream analysis), 4) number of necessary RT replicates and 5) possible inhibition of the reaction.

To summarize, RT is still not fully understood reaction. It appears that eventually all reaction components have an influence on its outcome, but it is their significance that varies. RTase itself can partially account for observed discrepancies, however, development of new engineered enzymes minimizes its impact. Despite this, there is not a single best performing RTase that is suitable for all applications. Nonetheless, RT can be considered a reproducible reaction, but one shall be aware of its downsides, especially when it is followed by methods of precise quantification. The reaction's weaknesses should be thoroughly further evaluated, with focus on the priming strategies and template annotation, leaving less uncertainty in our understanding.

9 References

Alvarez, M., Matamoros, T., & Menéndez-Arias, L. (2009). Increased thermostability and fidelity of DNA synthesis of wild-type and mutant HIV-1 group O reverse transcriptases. *J Mol Biol, 392*(4), 872-884.

Androvic, P., Valihrach, L., Elling, J., Sjoback, R., & Kubista, M. (2017). Two-tailed RT–qPCR: a novel method for highly accurate miRNA quantification. *Nucleic Acids Res, 45*(15), e144.

Arezi, B., & Hogrefe, H. (2009). Novel mutations in Moloney Murine Leukemia Virus reverse transcriptase increase thermostability through tighter binding to template-primer. *Nucleic Acids Res, 37*(2), 473-481.

Arezi, B., McCarthy, M., & Hogrefe, H. (2010). Mutant of Moloney murine leukemia virus reverse transcriptase exhibits higher resistance to common RT–qPCR inhibitors. *Anal Biochem, 400*(2), 301-303.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature, 226*(5252), 1209-1211.

Baltimore, D., Huang, A. S., & Stampfer, M. (1970). Ribonucleic acid synthesis of vesicular stomatitis virus, II. An RNA polymerase in the virion. *Proc Natl Acad Sci U S A, 66*(2), 572-576.

Bar, T., Kubista, M., & Tichopad, A. (2012). Validation of kinetics similarity in qPCR. *Nucleic Acids Res, 40*(4), 1395-1406.

Baranauskas, A., Paliksa, S., Alzbutas, G., Vaitkevicius, M., Lubiene, J., Letukiene, V., et al. (2012). Generation and characterization of new highly thermostable and processive M-MuLV reverse transcriptase variants. *Protein Eng Des Sel, 25*(10), 657-668.

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., et al. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science, 220*(4599), 868-871.

Beilhartz, G. L., & Götte, M. (2010). HIV-1 Ribonuclease H: Structure, Catalytic Mechanism and Inhibitors. *Viruses, 2*(4), 900-926.

Bengtsson, M., Karlsson, H. J., Westman, G., & Kubista, M. (2003). A new minor groove binding asymmetric cyanine reporter dye for real-time PCR. *Nucleic Acids Res, 31*(8), e45.

Bittner, J. J. (1936). Some possible effects of nursing on the mammary gland tumor Incidence in mice. *Science, 84*(2172), 162.

Boettiger, D., & Temin, H. M. (1970). Light inactivation of focus formation by chicken embryo fibroblasts infected with avian sarcoma virus in the presence of 5-bromodeoxyuridine. *Nature, 228*(5272), 622-624.

Brooks, E. M., Sheflin, L. G., & Spaulding, S. W. (1995). Secondary structure in the 3' UTR of EGF and the choice of reverse transcriptases affect the detection of message diversity by RT-PCR. *Biotechniques, 19*(5), 806-812, 814-805.

Bustin, S., Dhillon, H. S., Kirvell, S., Greenwood, C., Parker, M., Shipley, G. L., et al. (2015). Variability of the reverse transcription step: practical implications. *Clin Chem, 61*(1), 202-212.

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., et al. (2009). The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem, 55*(4), 611-622.

Caplin, B. E., Rasmussen, R. P., Bernard, P. S., & Wittwer, C. T. (1999). LightCycler™ Hybridization Probes - The most direct way to monitor PCR amplification for quantification and mutation detection. *Biochemica, 1*, 5-8.

Coté, M. L., & Roth, M. J. (2008). Murine leukemia virus reverse transcriptase: structural comparison with HIV-1 reverse transcriptase. *Virus Res, 134*(1-2), 186-202.

Crick, F. (1970). Central dogma of molecular biology. *Nature, 227*(5258), 561-563.

Crick, F. H. (1958). On protein synthesis. *Symp Soc Exp Biol, 12*, 138-163.

Das, D., & Georgiadis, M. M. (2004). The crystal structure of the monomeric reverse transcriptase from Moloney murine leukemia virus. *Structure, 12*(5), 819-829.

Devonshire, A. S., Elaswarapu, R., & Foy, C. A. (2010). Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. *BMC Genomics, 11*, 662.

Dickson, K. A., Haigis, M. C., & Raines, R. T. (2005). Ribonuclease inhibitor: structure and function. *Prog Nucleic Acid Res Mol Biol, 80*, 349-374.

Faridani, O. R., Abdullayev, I., Hagemann-Jensen, M., Schell, J. P., Lanner, F., & Sandberg, R. (2016). Single-cell sequencing of the small-RNA transcriptome. *Nat Biotechnol, 34*(12), 1264-1266.

Gibson, U. E., Heid, C. A., & Williams, P. M. (1996). A novel method for real time quantitative RT-PCR. *Genome Res, 6*(10), 995-1001.

Gilboa, E., Mitra, S. W., Goff, S., & Baltimore, D. (1979). A detailed model of reverse transcription and tests of crucial aspects. *Cell, 18*(1), 93-100.

Goldschmidt, V., Didierjean, J., Ehresmann, B., Ehresmann, C., Isel, C., & Marquet, R. (2006). Mg2+ dependency of HIV-1 reverse transcription, inhibition by nucleoside analogues and resistance. *Nucleic Acids Res, 34*(1), 42-52.

Gross, L. (1957). Filterable agent causing leukemia following inoculation into newborn mice. *Tex Rep Biol Med, 15*(3), 603-616; discussion 616-626.

Hickman, S. E., Kingery, N. D., Ohsumi, T. K., Borowsky, M. L., Wang, L. C., Means, T. K., et al. (2013). The microglial sensome revealed by direct RNA sequencing. *Nat Neurosci, 16*(12), 1896-1905.

Holland, P. M., Abramson, R. D., Watson, R., & Gelfand, D. H. (1991). Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A, 88*(16), 7276-7280.

Huang, H., Chopra, R., Verdine, G. L., & Harrison, S. C. (1998). Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science, 282*(5394), 1669-1675.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science, 343*(6172), 776-779.

Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., et al. (2006). The real-time polymerase chain reaction. *Mol Aspects Med, 27*(2-3), 95-125.

Kuo, K. W., Leung, M. F., & Leung, W. C. (1997). Intrinsic secondary structure of human TNFR-I mRNA influences the determination of gene expression by RT-PCR. *Mol Cell Biochem, 177*(1-2), 1-6.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921.

Lekanne Deprez, R. H., Fijnvandraat, A. C., Ruijter, J. M., & Moorman, A. F. (2002). Sensitivity and accuracy of quantitative real-time polymerase chain reaction using SYBR green I depends on cDNA synthesis conditions. *Anal Biochem, 307*(1), 63-69.

Levesque-Sergerie, J. P., Duquette, M., Thibault, C., Delbecchi, L., & Bissonnette, N. (2007). Detection limits of several commercial reverse transcriptase enzymes: impact on the low- and high-abundance transcript levels assessed by quantitative RT-PCR. *BMC Mol Biol, 8*, 93.

Lindén, J., Ranta, J., & Pohjanvirta, R. (2012). Bayesian modeling of reproducibility and robustness of RNA reverse transcription and quantitative real-time polymerase chain reaction. *Anal Biochem, 428*(1), 81-91.

Lingner, J., Hughes, T. R., Shevchenko, A., Mann, M., Lundblad, V., & Cech, T. R. (1997). Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science, 276*(5312), 561-567.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell, 161*(5), 1202-1214.

Malik, O., Khamis, H., Rudnizky, S., & Kaplan, A. (2017). The mechano-chemistry of a monomeric reverse transcriptase. *Nucleic Acids Res, 45*(22), 12954-12962.

Manganelli, R., Tyagi, S., & Smith, I. (2001). Real Time PCR Using Molecular Beacons : A New Tool to Identify Point Mutations and to Analyze Gene Expression in Mycobacterium tuberculosis. *Methods Mol Med, 54*, 295-310.

Meyer, P. R., Rutvisuttinunt, W., Matsuura, S. E., So, A. G., & Scott, W. A. (2007). Stable complexes formed by HIV-1 reverse transcriptase at distinct positions on the primer-template controlled by binding deoxynucleoside triphosphates or foscarnet. *J Mol Biol, 369*(1), 41-54.

Miranda, J. A., & Steward, G. F. (2017). Variables influencing the efficiency and interpretation of reverse transcription quantitative PCR (RT–qPCR): An empirical study using Bacteriophage MS2. *J Virol Methods, 241*, 1-10.

Mizuno, M., Yasukawa, K., & Inouye, K. (2010). Insight into the mechanism of the stabilization of moloney murine leukaemia virus reverse transcriptase by eliminating RNase H activity. *Biosci Biotechnol Biochem, 74*(2), 440-442.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods, 5*(7), 621-628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science, 320*(5881), 1344-1349.

Nowotny, M., Gaidamakov, S. A., Crouch, R. J., & Yang, W. (2005). Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell, 121*(7), 1005-1016.

Opel, K. L., Chung, D., & McCord, B. R. (2010). A study of PCR inhibition mechanisms using real time PCR. *J Forensic Sci, 55*(1), 25-33.

Oppermann, H., Levinson, A. D., Varmus, H. E., Levintow, L., & Bishop, J. M. (1979). Uninfected vertebrate cells contain a protein that is closely related to the product of the avian sarcoma virus transforming gene (src). *Proc Natl Acad Sci U S A, 76*(4), 1804-1808.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., et al. (2009). Direct RNA sequencing. *Nature, 461*(7265), 814-818.

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science, 344*(6190), 1396-1401.

Perk, K., & Moloney, J. B. (1966). Pathogenesis of a virus-induced rhabdomyosarcoma in mice. *J Natl Cancer Inst, 37*(5), 581-599.

Roberts, J. D., Bebenek, K., & Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science, 242*(4882), 1171-1173.

Rossen, L., Nørskov, P., Holmstrøm, K., & Rasmussen, O. F. (1992). Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions. *Int J Food Microbiol, 17*(1), 37-45.

Rous, P. (1911). A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *J Exp Med, 13*(4), 397-411.

Sarafianos, S. G., Das, K., Tantillo, C., Clark, A. D., Ding, J., Whitcomb, J. M., et al. (2001). Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J, 20*(6), 1449-1461.

Sarafianos, S. G., Marchand, B., Das, K., Himmel, D. M., Parniak, M. A., Hughes, S. H., et al. (2009). Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J Mol Biol, 385*(3), 693-713.

Schmidt, W. M., & Mueller, M. W. (1999). CapSelect: a highly sensitive method for 5' CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res, 27*(21), e31.

Singer, M. F. (1982). SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell, 28*(3), 433-434.

Stangegaard, M., Dufva, I. H., & Dufva, M. (2006). Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA. *Biotechniques, 40*(5), 649-657.

Ståhlberg, A., Håkansson, J., Xian, X., Semb, H., & Kubista, M. (2004b). Properties of the reverse transcription reaction in mRNA quantification. *Clin Chem, 50*(3), 509-515.

Ståhlberg, A., Kubista, M., & Pfaffl, M. (2004a). Comparison of reverse transcriptases in gene expression analysis. *Clin Chem, 50*(9), 1678-1680.

Suo, Z., & Johnson, K. A. (1998). DNA secondary structure effects on DNA synthesis catalyzed by HIV-1 reverse transcriptase. *J Biol Chem, 273*(42), 27259-27267.

Svensson, V., Natarajan, K. N., Ly, L. H., Miragaia, R. J., Labalette, C., Macaulay, I. C., et al. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat Methods, 14*(4), 381-387.

Temin, H. M. (1960). The control of cellular morphology in embryonic cells infected with rous sarcoma virus in vitro. *Virology, 10*, 182-197.

Temin, H. M. (1963). The effects of Actinomycin D on growth of Rous sarcoma virus in vitro. *Virology, 20*, 577-582.

Temin, H. M. (1964). The participation of DNA in Rous sarcoma virus production. *Virology, 23*, 486-494.

Temin, H. M., & Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature, 226*(5252), 1211-1213.

Valladares, Y. (1960). Studies on cancerous pathogenesis. Production of leukemia and polycythemia vera by means of cancerous nucleoproteins from tissue cultures. *Med Exp Int J Exp Med, 2*, 309-316.

Verman, I. M., Temple, G. F., Fan, H., & Baltimore, D. (1974). Synthesis by reverse transcriptase of DNA complementary to globin messenger RNA. *Basic Life Sci, 3*, 355-372.

Vermeulen, J., De Preter, K., Lefever, S., Nuytens, J., De Vloed, F., Derveaux, S., et al. (2011). Measurable impact of RNA quality on gene expression results from quantitative PCR. *Nucleic Acids Res, 39*(9), e63.

Wilson, I. G. (1997). Inhibition and facilitation of nucleic acid amplification. *Appl Environ Microbiol, 63*(10), 3741-3751.

Wu, W., Henderson, L. E., Copeland, T. D., Gorelick, R. J., Bosche, W. J., Rein, A., et al. (1996). Human immunodeficiency virus type 1 nucleocapsid protein reduces reverse transcriptase pausing at a secondary structure near the murine leukemia virus polypurine tract. *J Virol, 70*(10), 7132-7142.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol, 20*(9), 1131-1139.

Yu, M., Ting, D. T., Stott, S. L., Wittner, B. S., Ozsolak, F., Paul, S., et al. (2012). RNA sequencing of pancreatic circulating tumour cells implicates WNT signalling in metastasis. *Nature, 487*(7408), 510-513.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science, 347*(6226), 1138-1142.

Zeka, F., Vanderheyden, K., De Smet, E., Cuvelier, C. A., Mestdagh, P., & Vandesompele, J. (2016). Straightforward and sensitive RT–qPCR based gene expression analysis of FFPE samples. *Sci Rep, 6*, 21418.

Zhang, J., & Byrne, C. D. (1999). Differential priming of RNA templates during cDNA synthesis markedly affects both accuracy and reproducibility of quantitative competitive reverse-transcriptase PCR. *Biochem J, 337 (Pt 2)*, 231-241.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., & Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques, 30*(4), 892-897.